ESTIMATION AND INFERENCE IN SPARSE MULTIVARIATE REGRESSION AND CONDITIONAL GAUSSIAN GRAPHICAL MODELS UNDER AN UNBALANCED DISTRIBUTED SETTING

Ensiyeh Nezakati, Eugen Pircalabelu

LIDAM Discussion Paper ISBA
2023 / 21

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

Estimation and inference in sparse multivariate regression and conditional Gaussian graphical models under an unbalanced distributed setting

Ensiyeh Nezakati
and
Eugen Pircalabelu
Institute of Statistics, Biostatistics and Actuarial Sciences
UCLouvain
Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium

Abstract

This paper proposes a distributed estimation and inferential framework for sparse multivariate regression and conditional Gaussian graphical models under the unbalanced splitting setting. This type of data splitting arises when the datasets from different sources cannot be aggregated on one single machine or when the available machines are of different powers. In this paper, the number of covariates, responses and machines grow with the sample size, while sparsity is imposed. Debiased estimators of the coefficient matrix and of the precision matrix are proposed on every single machine and theoretical guarantees are provided. Moreover, new aggregated estimators that pool information across the machines using a pseudo log-likelihood function are proposed. It is shown that they enjoy efficiency and asymptotic normality as the number of machines grows with the sample size. The performance of these estimators is investigated via a simulation study

and a real data example. It is shown empirically that the performances of these estimators are close to those of the non-distributed estimators which use the entire dataset.

Keywords: Multivariate regression models; Conditional Gaussian graphical models; Debiased estimation; Precision matrix; Sparsity; Unbalanced distributed setting

1 Introduction

In multivariate linear regression, multiple response variables (say p) are regressed on multiple predictors (say q). This model has many applications in the real world, especially in genomic data analysis, where one can model the dependence of RNA levels on DNA copy numbers through a multivariate regression model with RNA levels being responses and the DNA copy numbers being predictors as in Peng et al. (2010). Moreover, a natural way to characterize the regularization network between a set of genetic variants and a set of expressions is through multivariate regression models (see for example, Wang et al., 2015; Yin and Li, 2011, 2013). Estimation of the coefficient matrix in multivariate regression models requires one to estimate pq coefficients, which becomes challenging with high-dimensional predictors and responses. Moreover, by adjusting the effect of covariates on the mean of the random response, we are able to estimate the structure of a conditional graphical model constructed using the elements of the precision matrix. Most studies focusing on the estimation of the precision matrix for Gaussian graphical models assume that the random vector has zero or constant mean. For a treatment on the subject in the high-dimensional context for a mean zero Gaussian graphical model, we refer the reader to Banerjee et al. (2008), Cai et al. (2011), Cai et al. (2016), Friedman et al. (2008), Loh and Tan (2018), Meinshausen and Bühlmann (2006), Ravikumar et al. (2011) and Yuan and Lin (2007) among many others. However, in many real applications, adjusting for the effect of covariates on the mean of the random vector is important for understanding the underlying graph structure.

Several studies used regularization-based approaches to estimate the coefficient and precision matrices with adjusted covariates. The works of Obozinski et al. (2011) and Yin and Li (2011) proposed a joint regularization penalty to estimate both the multivariate regression coefficients corresponding to the covariates and the precision matrix of a Gaussian graphical models iteratively.

In Rothman et al. (2010) and Wang (2015), the coefficient and precision matrices were estimated simultaneously via a joint penalized likelihood function. The works of Cai et al. (2013) and Yin and Li (2013) proposed a two-stage strategy which first estimates the regression coefficients and then using the residuals from the first stage, estimates the precision matrix. In Chen et al. (2016), a tuning parameter free estimator was proposed that is asymptotically normal and efficient for the estimation of every finite sub-graph for covariate adjusted Gaussian graphical models. In these studies, due to the use of penalization, the estimators are biased. In this paper, by introducing debiased estimators, we are able to perform not only the estimation, but also inference and hypothesis testing.

With the development of technology, the size of the datasets grows at a high rate, such that in certain situations it is not possible to store all needed datasets in the memory of one single machine. Moreover, in recent frameworks, like federated learning McMahan et al. (2017), due to privacy concerns, it may be impossible to collect datasets from different resources on one single central machine. As such, the dataset is partitioned onto a cluster of machines. Distributed statistical approaches, also known as 'divide and conquer' approaches, have drawn a lot of attention in the last decade and have been developed for various statistical problems. The two most popular techniques in distributed statistical inference and estimation problems are 'averaging' estimators from local machines and the 'one-step' approach, which combines the simple averaging estimator with a classical Newton's method to generate a one-step estimator. In Jordan et al. (2018), the authors presented a 'Communication-efficient Surrogate Likelihood' framework for solving distributed statistical inference problems in low-dimensional, high-dimensional and Bayesian frameworks. In Battey et al. (2018), Chen and Xie (2014) and Lee et al. (2017), the authors considered a high-dimensional sparse parameter vector estimation problem, where they adopted the penalized M-estimator setting. For a detailed review on aggregation methods for distributed estimators, the reader can refer to Huo and Cao (2019).

Nevertheless, most studies in the distributed setting have focused on the balanced sub-samples case, while in recent approaches like federated learning, some of the machines are more powerful than others and it is not efficient to distribute a dataset on different machines with equal sizes. In this situation, just taking a simple average is not an optimal approach for aggregating estimators. Recently, Nezakati and Pircalabelu (2023) proposed a new weighted, aggregated estimator for the elements of the precision matrix in a zero-mean

Gaussian graphical model, where the weights are a function of sub-sample sizes and the variances estimated based on the sub-samples. In this paper, we introduce new aggregated estimators for the coefficient and precision matrices in multivariate regression models using a pseudo log-likelihood function which is constructed using the asymptotic distribution of the debiased estimators. It is shown empirically that these estimators perform better than the simple average in terms of accuracy and coverage probabilities. These estimators are constructed for the setting where the number of responses (p), covariates (q) and machines (K) grow with the sample size (n). As a consequence, sparsity assumptions are imposed on the true matrices as a function of p, q and n. Different upper bounds are derived on the number of machines to guarantee the consistency and asymptotic normality of the estimators.

The content of this paper is organized as follows. Notation and preliminaries are presented in Section 2. The debiased distributed estimators and their statistical properties are provided for the coefficient matrix and the precision matrix separately in Sections 3 and 4, respectively. The final aggregated estimators for both target matrices are introduced in Section 5. Theoretical properties of these estimators are also investigated in this section. In Section 6, the performance of the estimators is evaluated by means of a controlled simulation study and in Section 7, the performance on a real data set is illustrated. We close with a discussion on the method in Section 8. Proofs of the supporting lemmas and theorems and more simulation results can be found in the Appendix.

2 Notation and preliminaries

The multivariate regression model in this paper is defined as

$$\dot{\mathbf{Y}} = \mathbf{\Gamma}^\top \dot{\mathbf{X}} + \dot{\varepsilon}, \tag{1}$$

where $\dot{\mathbf{Y}} = (Y^1, \ldots, Y^p)^\top \in \mathbb{R}^p$ and $\dot{\mathbf{X}} = (X^1, \ldots, X^q)^\top \in \mathbb{R}^q$ are the random response and covariate vectors, respectively. Moreover, the matrix $\mathbf{\Gamma}$ is a $q \times p$ regression coefficient matrix. Consider a random noise vector $\dot{\varepsilon} = (\varepsilon^1, \ldots, \varepsilon^p)^\top \in \mathbb{R}^p$ independent of $\dot{\mathbf{X}}$, which follows a p-dimensional Gaussian distribution with mean zero, covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. The components of $\dot{\mathbf{Y}}$ are mapped to the node set $\mathcal{V} = \{1, \ldots, p\}$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ describes the set of edges between all pairs $(a, b) \in \mathcal{V} \times \mathcal{V}$, $a \neq b$. An edge $(a, b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, is called

undirected if and only if both pairs (a, b) and (b, a) are included in $\mathcal{E}$. A pair (a, b) is included in the edge set $\mathcal{E}$ if and only if the variables Y^a and Y^b are conditionally dependent given $\dot{\mathbf{X}}$ and all remaining random variables in $\dot{\mathbf{Y}}$.

According to model (1), conditionally on $\dot{\mathbf{X}} = \dot{\mathbf{x}}$, $\dot{\mathbf{Y}}$ follows a p-dimensional Gaussian distribution with mean vector $\mathbf{\Gamma}^\top \dot{\mathbf{x}}$, covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Theta}$. Every off-diagonal entry (a, b) of $\mathbf{\Theta}$ is proportional to the partial correlation between Y^a and Y^b given $\dot{\mathbf{X}}$ and all other variables in $\dot{\mathbf{Y}}$. As such, a pair of variables in $\dot{\mathbf{Y}}$ is conditionally independent given all remaining variables of $\dot{\mathbf{Y}}$ and $\dot{\mathbf{X}}$, if and only if the corresponding entry in the precision matrix $\mathbf{\Theta}$ is zero. Denote $\mathbf{A}_{ab}$ as the (a, b)-th element of an arbitrary matrix $\mathbf{A}$. The support of the precision matrix $\mathbf{\Theta}$ is defined as the index set of its non-zero off-diagonal elements

$$S_1 := \{(a, b) \in \mathcal{V} \times \mathcal{V}, a \neq b \mid \mathbf{\Theta}_{ab} \neq 0\}$$

with cardinality $s_1 = \#S_1$ and the maximum node degree or row sparsity of $\mathbf{\Theta}$ is defined as

$$d_1 := \max_{a \in \mathcal{V}} \#\{b \in \mathcal{V}, b \neq a \mid \mathbf{\Theta}_{ab} \neq 0\}.$$

Analogously, the support of the regression coefficient matrix $\mathbf{\Gamma}$ is considered as

$$S_2 := \big\{(a, b) \in \{1, \ldots, q\} \times \{1, \ldots, p\} \mid \mathbf{\Gamma}_{ab} \neq 0\big\},$$

which is the index set of its non-zero elements, with cardinality $s_2 = \#S_2$. Moreover, the support of the b-th column of the regression coefficient matrix $\mathbf{\Gamma}$ is defined as $S_2(b) = \big\{a \in \{1, \ldots, q\} \mid \mathbf{\Gamma}_{ab} \neq 0\big\}$, $b = 1, \ldots, p$, with cardinality $s_2(b) = \#S_2(b)$.

Given n independent observations from the pair $(\dot{\mathbf{Y}}^\top, \dot{\mathbf{X}}^\top)$, our goal is to introduce distributed, debiased estimators for $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ from model (1). Suppose that the n samples are randomly divided into K non-overlapping unbalanced sub-samples with size n_k for the k-th sub-sample, $k = 1, \ldots, K$, and consider $n_\dagger = \min_{1 \leq k \leq K} n_k$, while $n = \sum_{k=1}^K n_k$. In this paper, p and q can grow with n, such that they might be larger than n. However, we suppose that $\log(p) = o(n_\dagger)$ and $\log(q) = o(\sqrt{n_\dagger})$, where $o(\cdot)$ expresses the asymptotic behavior of a sequence and is defined below. The usual assumption in the high-dimensional context is that the underlying matrices (coefficient and precision matrices in this paper) are sparse, which means that the number of non-zero elements in the matrices cannot grow too fast and a certain

bound is imposed on it. This sparsity reflects that many predictors in the regression model are redundant and that the graph related to the precision matrix has a rather low number of edges. To impose this sparsity condition on the estimation, one common approach is to add an ℓ_1 penalty to the function which is going to be optimized. This kind of regularization effectively forces some of the elements to zero, thus resulting in sparse solutions. There exists a wide variety of methods making use of ℓ_1 regularization, see for example Friedman et al. (2008), Ravikumar et al. (2011) and van de Geer et al. (2014). For convenience of notation, denote by $\mathbf{Y}_k = [\dot{\mathbf{Y}}_{1,k}, \ldots, \dot{\mathbf{Y}}_{n_k,k}]^\top$ and $\xi_k = [\dot{\varepsilon}_{1,k}, \ldots, \dot{\varepsilon}_{n_k,k}]^\top$ the k-th sub-sample of the response vector $\dot{\mathbf{Y}}$ and the random noise $\dot{\varepsilon}$, respectively, both arranged as matrices of dimension $n_k \times p$, with $\dot{\mathbf{Y}}_{l,k} \in \mathbb{R}^p$, $\dot{\varepsilon}_{l,k} \in \mathbb{R}^p$ as the l-th row, $l = 1, \ldots, n_k$, and by $\mathbf{X}_k = [\dot{\mathbf{X}}_{1,k}, \ldots, \dot{\mathbf{X}}_{n_k,k}]^\top$ the k-th design matrix of dimension $n_k \times q$, with $\dot{\mathbf{X}}_{l,k} \in \mathbb{R}^q$ as the l-th row, $l = 1, \ldots, n_k$. With this notation, the sample version of the regression model (1) corresponds to

$$\mathbf{Y}_k = \mathbf{X}_k \mathbf{\Gamma} + \xi_k, \tag{2}$$

where the rows of ξ_k are i.i.d. p-dimensional Gaussian vectors with mean zero, covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Theta}$. In the next two sections, debiased estimators for $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ based on the k-th sub-sample are derived.

Before starting the discussion, we introduce some order notation which is needed later in the paper. For two sequences $\{a_n; n \geq 1\}$ and $\{b_n; n \geq 1\}$, $b_n = O(a_n)$ if there exist positive numbers M_0 and N_0 such that $|b_n/a_n| \leq M_0$ for all $n \geq N_0$. Similarly, for a random sequence $\{X_n; n \geq 1\}$, we write $X_n = O_p(a_n)$ if for every $\epsilon > 0$, there exist finite numbers $M_0 > 0$ and $N_0 > 0$ such that $\mathbb{P}(|X_n/a_n| > M_0) < \epsilon$ for all $n \geq N_0$. We write $b_n \asymp a_n$ if both $b_n = O(a_n)$ and $a_n = O(b_n)$ hold. Moreover, $b_n = o(a_n)$ if $\lim_{n\to\infty} b_n/a_n = 0$. In the case of a random sequence $\{X_n; n \geq 1\}$, we write $X_n = o_p(a_n)$ if $X_n/a_n \xrightarrow{p} 0$, as $n \to \infty$, where the notation $\xrightarrow{p}$ denotes convergence in probability. For a matrix $\mathbf{A}$, we use the notation $|||\mathbf{A}|||_\infty = \max_a \sum_b |\mathbf{A}_{ab}|$ and $||\mathbf{A}||_\infty = \max_{a,b} |\mathbf{A}_{ab}|$ for the matrix and elementwise ℓ_∞ norms, respectively. The same symbol $||\mathbf{x}||_\infty = \max_b |\mathbf{x}_b|$ is used for the ℓ_∞ norm of a vector $\mathbf{x}$, where $\mathbf{x}_b$ is the b-th element of $\mathbf{x}$. Moreover, $||\mathbf{A}||_1 = \sum_{a,b} |\mathbf{A}_{ab}|$ and $||\mathbf{x}||_1 = \sum_b |\mathbf{x}_b|$ are used for the elementwise ℓ_1 norm of a matrix $\mathbf{A}$ and of a vector $\mathbf{x}$, respectively. We use $||\mathbf{A}||_F = \sqrt{\sum_{a,b} \mathbf{A}_{ab}^2} = \sqrt{\text{trace}(\mathbf{A}^\top \mathbf{A})}$ for the Frobenius norm of a matrix and $||\mathbf{x}||_2 = \sqrt{\sum_b \mathbf{x}_b^2}$ for the ℓ_2 norm of a vector $\mathbf{x}$. Finally, by $\mathbf{A} \otimes \mathbf{B}$ we denote the Kronecker product

of two arbitrary matrices $\mathbf{A}$ and $\mathbf{B}$ of dimension $m \times n$ and $p \times q$ as a $pm \times qn$ block matrix with $\mathbf{A}_{ab}\mathbf{B}$ for the block (a, b) where $\mathbf{A}_{ab}$ is the (a, b)-th element of matrix $\mathbf{A}$, $a = 1, \ldots, m$, and $b = 1, \ldots, n$.

3 Distributed estimator of the coefficient matrix using the k-th sub-sample

We make the following assumptions in this section.

(A1) The rows of the design matrix $\mathbf{X}_k$ are i.i.d. q-dimensional Gaussian vectors with mean zero and positive definite covariance matrix $\mathbf{Q}$, where $\max_a \mathbf{Q}_{aa} = O(1)$.

(A2) The eigenvalues of $\mathbf{Q}$ are bounded, i.e., there exists a constant $\Lambda_1 \geq 1$ such that $1/\Lambda_1 \leq \Lambda_{\min}(\mathbf{Q}) \leq \Lambda_{\max}(\mathbf{Q}) \leq \Lambda_1$, where $\Lambda_{\min}(\mathbf{Q})$ and $\Lambda_{\max}(\mathbf{Q})$ are the minimum and maximum eigenvalues of $\mathbf{Q}$, respectively.

Denote the maximum row sparsity of the inverse covariance matrix $\mathbf{Q}^{-1}$ with $d_2 := \max_{a \in \{1,\ldots,q\}} \#\{a' \in \{1, \ldots, q\}, a' \neq a \mid \mathbf{Q}^{-1}_{aa'} \neq 0\}$. To find an initial estimator for the regression coefficient matrix $\mathbf{\Gamma}$ in the k-th sub-sample, following Yin and Li (2013), we minimize the following joint penalized residual sum of squares and denote by $\hat{\mathbf{\Gamma}}_k$,

$$\hat{\mathbf{\Gamma}}_k = \arg\min_{\mathbf{\Gamma}} \left\{ \frac{1}{2n_k} \text{trace}\big\{(\mathbf{Y}_k - \mathbf{X}_k\mathbf{\Gamma})^\top (\mathbf{Y}_k - \mathbf{X}_k\mathbf{\Gamma})\big\} + \rho_k ||\mathbf{\Gamma}||_1 \right\}, \tag{3}$$

where $\rho_k > 0$ is a regularization parameter that forces $\hat{\mathbf{\Gamma}}_k$ to be sparse. Lemma 4 in Appendix B, provides a bound on the ℓ_1 norm of the difference between $\hat{\mathbf{\Gamma}}_k$ and the true matrix $\mathbf{\Gamma}$ under additional regularity conditions.

Due to the ℓ_1 penalty which is added to this loss function, $\hat{\mathbf{\Gamma}}_k$ is a biased estimator. To obtain a debiased estimator of $\mathbf{\Gamma}$, we use the idea of van de Geer et al. (2014) by inverting the Karush-Kuhn-Tucker (KKT) conditions. They showcased this method in the linear regression and generalized linear models framework. Using the KKT conditions in (3), we have

$$-\mathbf{X}_k^\top \mathbf{Y}_k / n_k + \mathbf{X}_k^\top \mathbf{X}_k \hat{\mathbf{\Gamma}}_k / n_k + \rho_k \mathbf{B}_k = \mathbf{0}, \tag{4}$$

where $\mathbf{B}_k$ belongs to the sub-differential of the ℓ_1 norm evaluated at $\hat{\mathbf{\Gamma}}_k$. By adding and subtracting $\mathbf{X}_k^\top \xi_k / n_k$ to (4), performing some algebra calculations

and finally vectorizing, we get

$$\sqrt{n_k}\text{vec}\big(\hat{\mathbf{\Gamma}}_k^d - \mathbf{\Gamma}\big) = \mathbf{T}_k + \mathbf{R}_{k,\mathbf{\Gamma}}, \tag{5}$$

where $\text{vec}(\cdot)$ is an operator which converts the matrix into a column vector by stacking the columns of the matrix on top of one another, $\hat{\mathbf{\Gamma}}_k^d = \hat{\mathbf{\Gamma}}_k + \mathbf{M}_k\mathbf{X}_k^\top(\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)/n_k$, $\mathbf{T}_k = \text{vec}\big(\mathbf{M}_k\mathbf{X}_k^\top\xi_k\big)/\sqrt{n_k}$, $\mathbf{R}_{k,\mathbf{\Gamma}} = \sqrt{n_k}\text{vec}\big((\mathbf{I}_q - \mathbf{M}_k\mathbf{C}_k)(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})\big)$, $\mathbf{I}_q$ is the identity matrix of dimension $q \times q$ and $\mathbf{M}_k$ plays a similar role to an inverse matrix. As $\mathbf{C}_k = \mathbf{X}_k^\top\mathbf{X}_k/n_k$, i.e. the sample covariance matrix using the sub-sample $\mathbf{X}_k$, is not always invertible, we consider $\mathbf{M}_k$ such that $\mathbf{M}_k\mathbf{C}_k \approx \mathbf{I}_q$. The estimator in (5) is the multivariate version of the one introduced in van de Geer et al. (2014). Note that the quantities in (5) are indexed by k, and as such one obtains a collection of debiased estimators $\hat{\mathbf{\Gamma}}_1^d, \ldots, \hat{\mathbf{\Gamma}}_K^d$, each using a particular sub-sample. In Section 5, we propose a novel, aggregated estimator based on the collection $\hat{\mathbf{\Gamma}}_1^d, \ldots, \hat{\mathbf{\Gamma}}_K^d$. In order to construct the aggregated estimator one needs the asymptotic distribution of $\hat{\mathbf{\Gamma}}_k^d$, $k = 1, \ldots, K$, which is investigated in the sequel.

Similarly to the case of a univariate response, by conditioning on $\mathbf{X}_k$, one can show the normality of $\mathbf{T}_k$ from (5). One just needs next to show that the remainder term $\mathbf{R}_{k,\mathbf{\Gamma}}$ vanishes with increasing n_k. To this end, we consider a suitable approximation for $\mathbf{M}_k$, such that with increasing n_k, the entries in $(\mathbf{I}_q - \mathbf{M}_k\mathbf{C}_k)$ get closer to zero. Several works (for instance, Javanmard and Montanari, 2018; van de Geer et al., 2014; Zhang and Zhang, 2014) assumed that $\mathbf{Q}^{-1}$ is sparse and then using the method of nodewise Lasso, estimated $\mathbf{Q}^{-1}$ and have set $\mathbf{M}_k = \hat{\mathbf{Q}}^{-1}$, where $\hat{\mathbf{Q}}^{-1}$ is the nodewise Lasso estimator of $\mathbf{Q}^{-1}$. We follow the same procedure, and for the reader's convenience, a brief description of the method is provided in Appendix A.1. Furthermore, we need to control the randomness of the product between the noise matrix ξ_k and the design matrix $\mathbf{X}_k$ in the multivariate linear regression. Many studies, focusing on Lasso regression models, control the randomness of the noise by conditioning on an event of interest (see for instance, Bühlmann and Van De Geer, 2011; Javanmard and Montanari, 2018). Similarly, considering the constant $2\rho_0 \leq \rho_k$, recall that ρ_k is the regularization parameter in (3), we consider the event

$$\mathcal{F}_k(n_k, p, q) = \bigg\{ ||\text{vec}\big(\mathbf{X}_k^\top\xi_k\big)||_\infty/n_k \leq \rho_0 \bigg\}.$$

Lemma 2 in Appendix B shows that for a suitable value of ρ_0, the event $\mathcal{F}_k(n_k, p, q)$ happens with a large probability. To show the asymptotic nor-

mality of $\mathbf{T}_k$ from (5), we need to impose as well a sparsity condition on the inverse covariance matrix $\mathbf{Q}^{-1}$. The sparsity condition in van de Geer et al. (2014) is considered as $d_2 = o(n/\log(q))$. In this paper, we need to restrict the elements of $\mathbf{Q}^{-1}$ to a sparser regime such that the maximum row sparsity grows at the rate $d_2 = o(\sqrt{n_\dagger}/\log(q))$. With this sparsity condition, we show the asymptotic normality of the final aggregated estimator in Section 5 as this condition is needed in Lemma 7, where we show that under such an assumption two sequences are equivalent in probability. Moreover, due to the fact that both K and n_k, $k = 1, \ldots, K$, grow in the distributed case, we impose the sparsity condition $s_2 = o\big(n_\dagger^{\pi_1}/(\log(q)\log(pq))\big)$, $0 < \pi_1 \leq 1/2$ on $\mathbf{\Gamma}$ to guarantee the theoretical properties of the aggregated estimator.

Theorem 1. *Consider the regression model* (2) *for the k-th sub-sample with zero-mean Gaussian noise matrix ξ_k having covariance matrix $\mathbf{\Sigma}$ and random design matrix $\mathbf{X}_k$, which satisfies assumptions* (A1) *and* (A2) *with sparsity condition $d_2 = o(\sqrt{n_\dagger}/\log(q))$. On the event $\mathcal{F}_k(n_k, p, q)$, with regularization parameter $\rho_k \asymp \sqrt{\log(pq)/n_k}$ in* (3) *and $\tilde{\rho}_{j,k} \asymp \sqrt{\log(q)/n_k}$, $j = 1, \ldots, q$, from the nodewise Lasso procedure, defined in Appendix A.1, we have*

$$\mathbf{T}_k|\mathbf{X}_k \sim \mathcal{N}_{pq}\big(\mathbf{0}, \mathbf{\Sigma} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)\big), \quad ||\mathbf{R}_{k,\mathbf{\Gamma}}||_\infty = O_p(s_2\sqrt{\log(q)\log(pq)/n_k}), \tag{6}$$

and under the additional sparsity condition $s_2 = o\big(n_\dagger^{\pi_1}/(\log(q)\log(pq))\big)$, $0 < \pi_1 \leq 1/2$, we have $||\mathbf{R}_{k,\mathbf{\Gamma}}||_\infty = o_p(1)$.

The proof is given in Appendix A.1.

One can use Theorem 1 to construct asymptotic inferential tools for $\mathbf{\Gamma}$ based on the k-th sub-sample. However, the covariance matrix $\mathbf{\Sigma}$ is in general unknown and it can be replaced in practice by a consistent estimator. Lemma 5 in Appendix B shows that if the eigenvalues of $\mathbf{\Sigma}$ are uniformly bounded, then under the random design setting, with sparsity condition $s_2 = o\big(n_\dagger^{\pi_1}/(\log(q)\log(pq))\big)$, $0 < \pi_1 \leq 1/2$, the estimator $\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} = (\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)^\top(\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)/n_k$ is a consistent estimator for the covariance matrix $\mathbf{\Sigma}$ with convergence rate in ℓ_∞ norm of order $O_p(\max\{\sqrt{\log(p)/n_k}, s_2\log(pq)/n_k\})$.

Using (5), one can also obtain the convergence rate of the debiased estimator $\hat{\mathbf{\Gamma}}_k^d$. This bound is investigated in Theorem 2.

Theorem 2. *Consider the regression model* (2) *for the k-th sub-sample with design matrix* $\mathbf{X}_k$ *which satisfies assumptions* (A1) *and* (A2)*. On the event* $\mathcal{F}_k(n_k, p, q)$*, with regularization parameter* $\rho_k \asymp \sqrt{\log(pq)/n_k}$*, we have*

$$||\hat{\mathbf{\Gamma}}_k^d - \mathbf{\Gamma}||_\infty = O_p\big(\max\{\sqrt{d_2 \log(pq)/n_k}, s_2\sqrt{\log(q)\log(pq)}/n_k\}\big), \qquad (7)$$

where under the assumption $\log(p)/\log(q) = o(n_\dagger^{\pi_2})$*,* $0 < \pi_2 < 1/2$*, and the sparsity conditions* $d_2 = o(\sqrt{n_\dagger}/\log(q))$ *and* $s_2 = o\big(n_\dagger^{\pi_1}/(\log(q)\log(pq))\big)$*,* $0 < \pi_1 \leq 1/2$*, we obtain* $||\hat{\mathbf{\Gamma}}_k^d - \mathbf{\Gamma}||_\infty = o_p(1)$*.*

The proof is given in Appendix A.2. In the next section, the distributed estimation of $\mathbf{\Theta}$ based on the k-th sub-sample is provided.

4 Distributed estimator of the precision matrix using the k-th sub-sample

Given the design matrix $\mathbf{X}_k$, the following assumptions (similarly to Jankova and van de Geer, 2015; Ravikumar et al., 2011) are adapted to our context and are considered for providing theoretical guarantees in the estimation procedure of $\mathbf{\Theta}$.

(B1) The eigenvalues of the precision matrix $\mathbf{\Theta}$ are bounded, i.e., there exists a constant $\Lambda_2 \geq 1$ such that $1/\Lambda_2 \leq \Lambda_{\min}(\mathbf{\Theta}) \leq \Lambda_{\max}(\mathbf{\Theta}) \leq \Lambda_2$, where $\Lambda_{\min}(\mathbf{\Theta})$ and $\Lambda_{\max}(\mathbf{\Theta})$ are the minimum and maximum eigenvalues of $\mathbf{\Theta}$, respectively. Moreover, $\max_a \mathbf{\Theta}_{aa} = O(1)$, where $\mathbf{\Theta}_{aa}$ is the a-th diagonal element of $\mathbf{\Theta}$.

Recall that $\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} = (\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)^\top(\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k)/n_k$ and consider $\mathbf{H}$ as the Hessian of the negative log-likelihood function proportional to $l(\mathbf{\Theta}) = \text{trace}(\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\mathbf{\Theta}) - \log\det(\mathbf{\Theta})$ which can be shown to be $\mathbf{H} = \mathbf{\Sigma} \otimes \mathbf{\Sigma}$ (see Jankova and van de Geer, 2015). By definition, $\mathbf{H}$ is a $p^2 \times p^2$ matrix indexed by the pair of elements from the node set, such that $\mathbf{H} = [\mathbf{H}_{(a,b),(c,d)}]$, where $(a,b),(c,d) \in \mathcal{V} \times \mathcal{V}$. Let $\mathbb{S}_1 = \big\{S_1 \cup \{(1,1),(2,2),\ldots,(p,p)\}\big\}$ with cardinality $\int_1$, which is equal to $\int_1 = s_1 + p$, and denote its complement set with $\mathbb{S}_1^c$.

(B2) The irrepresentability condition holds for the true precision matrix $\mathbf{\Theta}$, i.e., there exists $\alpha_1 \in (0,1]$ such that $\max_{e\in\mathbb{S}_1^c} ||\mathbf{H}_{e\mathbb{S}_1}\big(\mathbf{H}_{\mathbb{S}_1\mathbb{S}_1}\big)^{-1}||_1 \leq$

$1-\alpha_1$, where $\mathbf{H}_{\mathbb{S}_1\mathbb{S}_1} \in \mathbb{R}^{\int_1 \times \int_1}$ is a sub-matrix of $\mathbf{H}$ whose rows and columns are indexed by the elements of $\mathbb{S}_1$. Moreover, e is a pair $(a,b) \in \mathbb{S}_1^c$ such that $\mathbf{H}_{e\mathbb{S}_1}$ is an $\int_1$-dimensional column vector with elements $\mathbf{H}_{e,(c,d)}$, where $(c,d) \in \mathbb{S}_1$.

This assumption is a necessary and sufficient condition for the graphical Lasso procedure to recover the edge-based (active) variables from the non-edge (non-active) ones.

Given $\mathbf{X}_k$, one can construct an estimator for $\mathbf{\Theta}$ using the following graphical Lasso optimization problem

$$\hat{\mathbf{\Theta}}_k = \arg\min_{\mathbf{\Theta} \in \mathcal{S}_{++}^p} \left\{ \text{trace}(\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \mathbf{\Theta}) - \log\det(\mathbf{\Theta}) + \lambda_k ||\mathbf{\Theta}||_{1,\text{off}} \right\}, \tag{8}$$

where $||\cdot||_{1,\text{off}}$ is the ℓ_1 off-diagonal norm of the matrix defined as $||\mathbf{\Theta}||_{1,\text{off}} = \sum_{a \neq b} |\mathbf{\Theta}_{ab}|$ and $\mathcal{S}_{++}^p$ is the space of positive definite matrices of dimension $p \times p$. To obtain a debiased estimator of $\mathbf{\Theta}$, similarly to Jankova and van de Geer (2015), by inverting the KKT conditions from the optimization problem (8), we have

$$\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \hat{\mathbf{\Theta}}_k^{-1} + \lambda_k \hat{\mathbf{D}}_k = \mathbf{0}, \tag{9}$$

where the matrix $\hat{\mathbf{D}}_k$ belongs to the sub-differential of the ℓ_1 off-diagonal norm evaluated at $\hat{\mathbf{\Theta}}_k$. The difference between (9) and the problem in Jankova and van de Geer (2015) is that the previous work used the sample covariance matrix $\mathbf{Y}_k^\top \mathbf{Y}_k / n_k$, but here due to the non-zero mean of the model, we use $\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}$ which contains the plug-in estimation of the regression coefficient matrix $\mathbf{\Gamma}$ therein, thus making the analysis more complicated. To simplify notation, define

$$\mathbf{W}_k = \hat{\mathbf{\Sigma}}_{k,\mathbf{\Gamma}} - \mathbf{\Sigma}, \quad \mathbf{W}_k^{'} = \hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}, \quad \mathbf{W}_k^{''} = \hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \hat{\mathbf{\Sigma}}_{k,\mathbf{\Gamma}},$$

where

$$\hat{\mathbf{\Sigma}}_{k,\mathbf{\Gamma}} = (\mathbf{Y}_k - \mathbf{X}_k\mathbf{\Gamma})^\top (\mathbf{Y}_k - \mathbf{X}_k\mathbf{\Gamma})/n_k.$$

Using the KKT condition (9) and performing some algebra calculations, leads to

$$\sqrt{n_k}(\hat{\mathbf{\Theta}}_k^d - \mathbf{\Theta}) = -\sqrt{n_k}\mathbf{\Theta}\mathbf{W}_k\mathbf{\Theta} + \mathbf{R}_{k,\mathbf{\Theta}}, \tag{10}$$

where $\hat{\mathbf{\Theta}}_k^d = 2\hat{\mathbf{\Theta}}_k - \hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\hat{\mathbf{\Theta}}_k$ is the k-th debiased estimator of $\mathbf{\Theta}$ and the term $\mathbf{R}_{k,\mathbf{\Theta}}$ is defined as

$$\mathbf{R}_{k,\mathbf{\Theta}} := -\sqrt{n_k}(\hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{I}_p)(\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}) - \sqrt{n_k}(\hat{\mathbf{\Theta}}_k - \mathbf{\Theta})\mathbf{W}_k^{'}\mathbf{\Theta} - \sqrt{n_k}\mathbf{\Theta}\mathbf{W}_k^{''}\mathbf{\Theta}. \tag{11}$$

In the sequel, it is shown that under suitable conditions, $||\mathbf{R}_{k,\mathbf{\Theta}}||_\infty = o_p(1)$ and that the term $\sqrt{n_k}\mathbf{\Theta}\mathbf{W}_k\mathbf{\Theta}$ is elementwise asymptotically normal.

Relative to the work of Jankova and van de Geer (2015), in addition to different covariance matrices used in (11), our remainder contains the extra term $\sqrt{n_k}\mathbf{\Theta}\mathbf{W}_k^{''}\mathbf{\Theta}$. This extra term is bounded in elementwise ℓ_∞ norm at the rate of $O_p(d_1 s_2 \log(pq)/\sqrt{n_k})$, as it is shown in Lemma 1. In the estimation procedure, if one considers $\mathbf{\Gamma} = \mathbf{0}$ as a special case, then the KKT condition (9) will simplify to the one in Jankova and van de Geer (2015), and as such the estimator we propose, $\hat{\mathbf{\Theta}}_k^d$, will simplify to the same debiased estimator with the same convergence rate from that work. Denote $\kappa_{\mathbf{\Sigma}} := |||\mathbf{\Sigma}|||_\infty$ as the matrix ℓ_∞ norm of $\mathbf{\Sigma}$ and $\kappa_{\mathbf{H}} = |||\big(\mathbf{H}_{\mathbb{S}_1,\mathbb{S}_1}\big)^{-1}|||_\infty$ as the matrix ℓ_∞ norm of the inverse of $\mathbf{H}_{\mathbb{S}_1,\mathbb{S}_1}$. In Lemma 1, negligibility of $\mathbf{R}_{k,\mathbf{\Theta}}$ is shown.

Lemma 1. Consider the multivariate regression model (2) with random Gaussian noise matrix ξ_k having zero-mean rows, covariance matrix $\mathbf{\Sigma}$ and precision matrix $\mathbf{\Theta}$. Let assumptions (B1)-(B2), (C1)-(C3) from Appendix A.3 and (46) from Appendix B hold. Consider the debiased estimator in (10) with regularization parameter $\lambda_k \asymp \sqrt{\log(p)/n_k}$. On the event $\mathcal{F}_k(n_k, p, q)$ with $2\rho_0 \leq \rho_k$, where $\rho_k \asymp \sqrt{\log(pq)/n_k}$, and under the assumptions $1/\alpha_1 = O(1)$, $\kappa_{\mathbf{\Sigma}} = O(1)$ and $\kappa_{\mathbf{H}} = O(1)$, we have

$$||\mathbf{R}_{k,\mathbf{\Theta}}||_\infty = O_p\bigg(\max\big\{d_1^{3/2}\log(p)/\sqrt{n_k}, d_1^2(\log(p))^{3/2}/n_k, d_1 s_2\log(pq)/\sqrt{n_k}\big\}\bigg), \tag{12}$$

and under the sparsity conditions $d_1^{3/2} = o\big(\sqrt{n_\dagger}/\log(p)\big)$ and $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, where $0 < \pi_3 \leq 1/6$, we get $||\mathbf{R}_{k,\mathbf{\Theta}}||_\infty = o_p(1)$.

The proof is given in Appendix A.3.

Remark 1. A similar convergence rate for the remainder term is obtained in Jankova and van de Geer (2015) but for the zero-mean model, which is of order $O_p\big(\max\big\{d_1^{3/2}\log(p)/\sqrt{n_k}, d_1^2(\log(p))^{3/2}/n_k\big\}\big)$. Comparing it with

(12), it is observed that the convergence rate of the extra term $\sqrt{n_k}\boldsymbol{\Theta}\mathbf{W}_k''\boldsymbol{\Theta}$ is of order $O_p(d_1 s_2 \log(pq)/\sqrt{n_k})$ which also adds more constraints on the negligibility of $\mathbf{R}_{k,\boldsymbol{\Theta}}$.

Theorem 3 investigates the elementwise asymptotic normality of the debiased estimator $\hat{\boldsymbol{\Theta}}_k^d$, which will be leveraged further in Section 5 to construct an aggregated estimator using all K estimators $\hat{\boldsymbol{\Theta}}_1^d, \ldots, \hat{\boldsymbol{\Theta}}_K^d$.

Theorem 3. *Under the assumptions of* Lemma 1 *and the sparsity conditions* $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$ *and* $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$, *for all* $(a, b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, *it holds that*

$$\sqrt{n_k}(\hat{\boldsymbol{\Theta}}_{ab,k}^d - \boldsymbol{\Theta}_{ab})/\sigma_{ab} = \mathbf{Z}_{ab,k} + o_p(1), \tag{13}$$

where $\boldsymbol{\Theta}_{ab}$ *and* $\hat{\boldsymbol{\Theta}}_{ab,k}^d$ *are the* (a, b)*-th element of* $\boldsymbol{\Theta}$ *and* $\hat{\boldsymbol{\Theta}}_k^d$, *respectively, and* $\sigma_{ab}^2 = \boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb} + \boldsymbol{\Theta}_{ab}^2$. *Moreover,* $1/\sigma_{ab} = O(1)$ *and* $\mathbf{Z}_{ab,k}$ *converges weakly to* $\mathcal{N}(0, 1)$ *as* n_k *grows.*

The proof is given in Appendix A.4.

Remark 2. One can use Theorem 3 to construct asymptotic confidence intervals and hypothesis testing procedures. However, to perform inference, one needs a consistent estimator for σ_{ab} as it is unknown. By similar arguments as in Lemma 2 of Jankova and van de Geer (2015) and considering $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$, the estimator $\hat{\sigma}_{ab,k}^2 = \hat{\boldsymbol{\Theta}}_{aa,k}\hat{\boldsymbol{\Theta}}_{bb,k} + \hat{\boldsymbol{\Theta}}_{ab,k}^2$ is a consistent estimator for σ_{ab}^2 with convergence rate $O_p(\max\{\log(p)/n_k, \sqrt{d_1 \log(p)/n_k}\})$.

Remark 3. To quantify the convergence rate of the debiased estimator $\hat{\boldsymbol{\Theta}}_k^d$, from (10), we have that

$$||\hat{\boldsymbol{\Theta}}_k^d - \boldsymbol{\Theta}||_\infty \leq |||\boldsymbol{\Theta}|||_\infty^2 ||\mathbf{W}_k||_\infty + ||\mathbf{R}_{k,\boldsymbol{\Theta}}||_\infty/\sqrt{n_k}.$$

Given $\mathbf{X}_k$ and under assumption (B1), by considering $\xi_k = \mathbf{Y}_k - \mathbf{X}_k\boldsymbol{\Gamma}$ and setting $\mathbf{A} = \mathbf{B} = \mathbf{I}_p$ in Lemma 2 of Lam and Fan (2009), for which the conditions are fulfilled, one can write $||\mathbf{W}_k||_\infty = O_p(\sqrt{\log(p)/n_k})$. Combining this bound with (12) and (36) from Appendix A.3, we get

$$||\hat{\boldsymbol{\Theta}}_k^d - \boldsymbol{\Theta}||_\infty = O_p\Big(\max\big\{ d_1\sqrt{\log(p)/n_k}, d_1^{3/2}\log(p)/n_k, \\ d_1^2(\log(p)/n_k)^{3/2}, d_1 s_2 \log(pq)/n_k \big\} \Big),$$

and under the sparsity conditions $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$, and $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$, the consistency of $\hat{\boldsymbol{\Theta}}_k^d$ follows.

5 The final aggregated estimators across the sub-samples

The results in the previous sections are provided at the level of the k-th sub-sample, $k = 1, \ldots, K$. To construct more reliable estimators, we aggregate estimators drawn from different distributed locations. There are multiple ways to aggregate these K estimators into a combined estimator. In this paper, due to the asymptotic normal distribution of the local estimators, we derive the final aggregated estimator by maximizing a pseudo log-likelihood function, which is constructed using the asymptotic normal density of local estimators constructed based on the sub-samples. Consider Δ as a general parameter of interest, for example $\mathbf{\Theta}_{ab}$ or $\text{vec}(\mathbf{\Gamma})_a$ in this paper, where $\text{vec}(\mathbf{\Gamma})_a$ is the a-th element of the vectorized form of $\mathbf{\Gamma}$, and $\hat{\Delta}_k$ as the k-th estimator based on the k-th sub-sample with variance σ^2 and denote its consistent estimator by $\hat{\sigma}_k^2$ which is also derived based on the k-th sub-sample. Consider the asymptotic normal density of $\hat{\Delta}_k$ at point ι_k as $\hat{f}_k(\iota_k \mid \Delta, \hat{\sigma}_k)$, where the variance σ^2 is replaced by its consistent estimator $\hat{\sigma}_k^2$. By maximizing the pseudo log-likelihood function constructed as

$$l(\Delta) = \log\left(\prod_{k=1}^{K} \hat{f}_k(\iota_k \mid \Delta, \hat{\sigma}_k)\right) \propto \sum_{k=1}^{K} (-n_k/2)(\iota_k - \Delta)^2/\hat{\sigma}_k^2,$$

with respect to Δ, the final aggregated estimator is of the form

$$\tilde{\Delta}_{\text{owAvg}} = \frac{1}{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}_k^2}} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}_k^2} \hat{\Delta}_k, \tag{14}$$

where the subscript "owAvg" stands for the optimally weighted average. We call this estimator an "optimally weighted average", as it is obtained using a maximization problem. Substituting debiased estimators $\hat{\mathbf{\Gamma}}_k^d$ and $\hat{\mathbf{\Theta}}_k^d$ in (14), one can aggregate distributed estimators of the coefficient and precision matrices from the sub-samples into the final combined estimators $\tilde{\mathbf{\Gamma}}_{\text{owAvg}}$ and $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$, respectively. Note that if the variance σ^2 is known, then replacing $\hat{\sigma}_k^2$ by σ^2 simplifies the aggregated estimator $\tilde{\Delta}_{\text{owAvg}}$ to two special cases. In the case of unbalanced sub-samples, $\tilde{\Delta}_{\text{owAvg}}$ is simplified to the sample size weighted average estimator

$$\tilde{\Delta}_{\text{wAvg}} = \sum_{k=1}^{K} \frac{n_k}{n} \hat{\Delta}_k, \tag{15}$$

where "wAvg" stands for the weighted average proportional to the sub-sample sizes. Similarly to the optimally weighted average estimator, by substituting debiased estimators $\hat{\mathbf{\Gamma}}_k^d$ and $\hat{\mathbf{\Theta}}_k^d$ in (15), the weighted average aggregated estimators of $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ can be constructed and denoted by $\tilde{\mathbf{\Gamma}}_{\text{wAvg}}$ and $\tilde{\mathbf{\Theta}}_{\text{wAvg}}$, respectively. Moreover, if the sub-samples are also balanced, $\tilde{\Delta}_{\text{owAvg}}$ is further simplified to the simple average estimator

$$\tilde{\Delta}_{\text{sAvg}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\Delta}_k, \tag{16}$$

where "sAvg" stands for the simple average. By the same argument as for the optimally weighted and weighted averages, the simple average estimators of $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ can be constructed and denoted by $\tilde{\mathbf{\Gamma}}_{\text{sAvg}}$ and $\tilde{\mathbf{\Theta}}_{\text{sAvg}}$, respectively. As an example of applying simple average to the penalized regression models, the reader can refer to Battey et al. (2018) which studied distributed parameter estimation methods for penalized regression and established the oracle asymptotic property of the averaged debiased estimators. However, from a finite sample perspective, the simple average estimator tends to underperform severely for unbalanced settings, as it will be shown in Section 6. The weighted average proportional to the sub-sample sizes is slightly more adapted as it accounts for the unbalanced setting by weighting according to the size of the samples on each machine. Note that as the variances of the estimators are unknown in practice, in the sequel, the special cases (15) and (16) are considered as general simple and weighted average estimators with unknown variances and their asymptotic distribution as well as the asymptotic distribution of the optimally weighted average estimator are investigated in Theorems 4 and 5 as $n_\dagger$ and K both grow, where we remind the reader that $n_\dagger = \min_{1\le k\le K} n_k$.

Using (5) and (14), by considering the consistent estimator $\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}$ for $\mathbf{\Sigma}$, the optimally weighted average estimator for the a-th element of $\text{vec}(\mathbf{\Gamma})$, $a = 1, \ldots, qp$, is of the form

$$\begin{aligned}\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a = &\left(\sum_{k=1}^{K}\frac{n_k}{\left[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)\right]_{aa}}\right)^{-1}\\ &\times\sum_{k=1}^{K}\frac{n_k}{\left[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)\right]_{aa}}\text{vec}\big(\hat{\mathbf{\Gamma}}_k^d\big)_a.\end{aligned} \tag{17}$$

It can be shown that

$$\sqrt{\sum_{k=1}^{K} n_k / [\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} \big(\mathrm{vec}(\tilde{\boldsymbol{\Gamma}}_{\mathrm{owAvg}})_a - \mathrm{vec}(\boldsymbol{\Gamma})_a\big) = \mathbf{W}_{a,\boldsymbol{\Gamma}} + \mathbf{R}_{a,\boldsymbol{\Gamma}}, \tag{18}$$

where

$$\mathbf{R}_{a,\boldsymbol{\Gamma}} = \sqrt{\frac{\sum_{k=1}^{K} n_k / [\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K} n_k / [\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}} \times \sum_{k=1}^{K} \frac{\sqrt{n_k [\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}}{\sqrt{n} [\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} \mathbf{R}_{a,k,\boldsymbol{\Gamma}},$$

$$\mathbf{W}_{a,\boldsymbol{\Gamma}} = \sqrt{\frac{\sum_{k=1}^{K} n_k / [\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K} n_k / [\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}} \times \sum_{k=1}^{K} \frac{\sqrt{n_k [\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}}{\sqrt{n} [\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} \mathbf{T}_{a,k},$$

and $\mathbf{R}_{a,k,\boldsymbol{\Gamma}}$ and $\mathbf{T}_{a,k}$ are the a-th element of $\mathbf{R}_{k,\boldsymbol{\Gamma}}$ and $\mathbf{T}_k$, respectively, defined in (5).

Similarly, using (10) and (14), by considering the consistent estimator $\hat{\sigma}^2_{ab,k}$ for σ^2_{ab} in (13), the aggregated estimator for the (a,b)-th element of $\boldsymbol{\Theta}$, $(a,b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, is of the form

$$\tilde{\boldsymbol{\Theta}}_{ab,\mathrm{owAvg}} = \left(\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}} \right)^{-1} \times \sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}} \hat{\boldsymbol{\Theta}}^d_{ab,k}. \tag{19}$$

Moreover,

$$\sqrt{\sum_{k=1}^{K} \frac{n_k}{\hat{\sigma}^2_{ab,k}}} \left(\tilde{\boldsymbol{\Theta}}_{ab,\mathrm{owAvg}} - \boldsymbol{\Theta}_{ab} \right) = \mathbf{W}_{ab,\boldsymbol{\Theta}} + \mathbf{R}_{ab,\boldsymbol{\Theta}}, \tag{20}$$

where

$$\mathbf{R}_{ab,\boldsymbol{\Theta}} = \sqrt{\frac{\sum_{k=1}^{K} n_k / \sigma^2_{ab}}{\sum_{k=1}^{K} n_k / \hat{\sigma}^2_{ab,k}}} \sum_{k=1}^{K} \frac{\sqrt{n_k} \sigma_{ab}}{\sqrt{n} \hat{\sigma}^2_{ab,k}} \mathbf{R}_{ab,k,\boldsymbol{\Theta}},$$

$$\mathbf{W}_{ab,\boldsymbol{\Theta}} = \sqrt{\frac{\sum_{k=1}^{K} n_k / \sigma^2_{ab}}{\sum_{k=1}^{K} n_k / \hat{\sigma}^2_{ab,k}}} \sum_{k=1}^{K} \frac{\sigma_{ab}}{\sqrt{n} \hat{\sigma}^2_{ab,k}} \sum_{l=1}^{n_k} \big(\boldsymbol{\Theta}_a^\top (\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k}) \times (\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k})^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab} \big),$$

and $\mathbf{R}_{ab,k,\boldsymbol{\Theta}}$ is the (a,b)-th element of $\mathbf{R}_{k,\boldsymbol{\Theta}}$ from (10), $\dot{\mathbf{Y}}_{l,k} \in \mathbb{R}^p$ and $\dot{\mathbf{X}}_{l,k} \in \mathbb{R}^q$ are the l-th row of $\mathbf{Y}_k$ and $\mathbf{X}_k$, respectively, while $\boldsymbol{\Theta}_a$ and $\boldsymbol{\Theta}_b$ are p-dimensional vectors coming from the a-th row and the b-th row of $\boldsymbol{\Theta}$, respectively.

Theorem 4. *Consider the regression model* (2)*, where* $\mathbf{X}_k$ *satisfies assumptions* (A1) *and* (A2)*, and consider the maximum row sparsity of* $\mathbf{Q}^{-1}$ *as* $d_2 = o(\sqrt{n_\dagger}/\log(q))$*. Moreover, consider the coefficient matrix* $\boldsymbol{\Gamma}$ *with sparsity condition* $s_2 = o\big(n_\dagger^{\pi_1}/(\log(pq)\log(q))\big)$*,* $0 < \pi_1 \leq 1/2$*. Suppose that* $\hat{\boldsymbol{\Gamma}}_k^d$*,* $k = 1, \ldots, K$*, is the* k*-th debiased estimator in* (5) *with tuning parameter* $\rho_k \asymp \sqrt{\log(pq)/n_k}$ *and let* $n_k/n \to c_k \in (0,1)$ *as* n_k *grows such that* $\lim_{K\to\infty} \sum_{k=1}^K c_k = 1$*. Then, on the event* $\mathcal{F}_k(n_k, p, q)$*,*

a) If K *grows at the rate of* $K = O\big(n^{\pi_4}/(\sqrt{\log(pq)\log(q)}\max\{s_2, \sqrt{d_2}\})\big)$*,* $0 < \pi_4 < 1/4$*, and* $\log(p)/\log(q) = o(n_\dagger^{\pi_2})$*,* $0 < \pi_2 < 1/2$*, for the* a*-th element of the vectorized form of* $\tilde{\boldsymbol{\Gamma}}_{\text{owAvg}}$*,* $a = 1, \ldots, qp$*, in* (18)*, we have*

$$\mathbf{W}_{a,\boldsymbol{\Gamma}} \xrightarrow{d} \mathcal{N}(0,1), \quad \textit{and} \quad |\mathbf{R}_{a,\boldsymbol{\Gamma}}| = o_p(1), \tag{21}$$

where $\xrightarrow{d}$ *denotes convergence in distribution.*

b) If K *grows at the rate of* $K = O\big(n^{\pi_5}/(\sqrt{\log(pq)\log(q)}\max\{s_2, d_2\})\big)$*,* $0 < \pi_5 < 1/2$*, in the spirit of the special case* (15)*, for the* a*-th element of the vectorized form of* $\tilde{\boldsymbol{\Gamma}}_{\text{wAvg}}$*,* $a = 1, \ldots, qp$*, which is denoted by* $\text{vec}(\tilde{\boldsymbol{\Gamma}}_{\text{wAvg}})_a$*, we have*

$$\frac{\sqrt{nK}}{\sqrt{\sum_{k=1}^K [\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}} \big(\text{vec}(\tilde{\boldsymbol{\Gamma}}_{\text{wAvg}})_a - \text{vec}(\boldsymbol{\Gamma})_a\big) = \mathbf{W}'_{a,\boldsymbol{\Gamma}} + \mathbf{R}'_{a,\boldsymbol{\Gamma}},$$

where $\mathbf{W}'_{a,\boldsymbol{\Gamma}}$ *converges weakly to* $\mathcal{N}(0,1)$ *and* $|\mathbf{R}'_{a,\boldsymbol{\Gamma}}| = o_p(1)$*.*

c) If K *grows at the rate of* $\sqrt{K} = O\big(n_\dagger^{\pi_6}/(\sqrt{\log(pq)\log(q)}\max\{s_2, d_2\})\big)$*,* $0 < \pi_6 < 1/2$*, in the spirit of the special case* (16)*, for the* a*-th element of the vectorized form of* $\tilde{\boldsymbol{\Gamma}}_{\text{sAvg}}$*,* $a = 1, \ldots, qp$*, which is denoted by* $\text{vec}(\tilde{\boldsymbol{\Gamma}}_{\text{sAvg}})_a$*, we have*

$$\frac{K^{3/2}}{\sqrt{(\sum_{k=1}^K [\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa})(\sum_{k=1}^K 1/n_k)}} \big(\text{vec}(\tilde{\boldsymbol{\Gamma}}_{\text{sAvg}})_a - \text{vec}(\boldsymbol{\Gamma})_a\big)$$
$$= \mathbf{W}''_{a,\boldsymbol{\Gamma}} + \mathbf{R}''_{a,\boldsymbol{\Gamma}},$$

where $\mathbf{W}''_{a,\mathbf{\Gamma}}$ *converges weakly to* $\mathcal{N}(0,1)$ *and* $|\mathbf{R}''_{a,\mathbf{\Gamma}}| = o_p(1)$.

The proof is given in Appendix A.5.

Remark 4. Note that Theorem 1 in Section 3 provides the asymptotic normality, conditionally on the design matrix $\mathbf{X}_k$, while the asymptotic normality result in Theorem 4 is more general as it is an unconditional result. By using a similar argument to the proof of Theorem 4, it can be shown that both $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a$ and $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{wAvg}})_a$ have an asymptotic variance equal to $[\mathbf{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}$ which implies that their asymptotic relative efficiency is equal to 1. This result is expected because (i) the definitions of $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a$ and $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{wAvg}})_a$ are closely related, and (ii) the estimated variances based on the sub-samples are consistent estimators for the true variance, i.e., $[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}/[\mathbf{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa} \xrightarrow{p} 1$. A comparison of these two estimators from a finite sample perspective is also provided through simulation and real data examples in Sections 6 and 7, respectively.

In Theorem 5, the asymptotic normality of the estimator in (19) and its two special cases is investigated as $n_\dagger$ and K both grow.

Theorem 5. *Consider the regression model* (2) *and suppose that assumptions* (B1)-(B2), (C1)-(C3) *from Appendix A.3 and* (46) *from Appendix B hold. Moreover, consider the coefficient matrix* $\mathbf{\Gamma}$ *with sparsity condition* $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$. *Suppose that* $\hat{\mathbf{\Theta}}_k^d$, $k = 1, \ldots, K$, *is the* k*-th debiased estimator in* (10) *with tuning parameter* $\lambda_k \asymp \sqrt{\log(p)/n_k}$ *and let* $n_k/n \to c_k \in (0,1)$ *as* n_k *grows, such that* $\lim_{K\to\infty} \sum_{k=1}^K c_k = 1$.

a) If K *and the maximum node degree of* $\mathbf{\Theta}$ *grow at the rates of* $K = O\big(n^{\pi_7}/(d_1 \log(p))\big)$, $0 < \pi_7 \leq 1/3$, *and* $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$, *respectively, then for every pair* $(a,b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, *of the pooled estimator* $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ *in* (20), *we have*

$$\mathbf{W}_{ab,\mathbf{\Theta}} \xrightarrow{d} \mathcal{N}(0,1), \quad \textit{and} \quad |\mathbf{R}_{ab,\mathbf{\Theta}}| = o_p(1). \tag{22}$$

b) Under the same conditions as in part a), in the spirit of the special case (15), *for every pair* $(a,b) \in \mathcal{V} \times \mathcal{V}$, $a \neq b$, *of* $\tilde{\mathbf{\Theta}}_{\text{wAvg}}$, *which is denoted by* $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}}$, *we have*

$$\frac{\sqrt{nK}}{\sqrt{\sum_{k=1}^K \hat{\sigma}_{ab,k}^2}}(\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}} - \mathbf{\Theta}_{ab}) = \mathbf{W}'_{ab,\mathbf{\Theta}} + \mathbf{R}'_{ab,\mathbf{\Theta}},$$

where $\mathbf{W}'_{ab,\mathbf{\Theta}}$ *converges weakly to* $\mathcal{N}(0,1)$ *and* $|\mathbf{R}'_{ab,\mathbf{\Theta}}| = o_p(1)$.

c) *If* K *and the maximum node degree of* $\mathbf{\Theta}$ *grow at the rates of* $K = O\big(n_\dagger^{1/3}/n^{1/4}\big)$ *and* $d_1^{3/2} = o\big(n_\dagger^{1/3}/\log(p)\big)$, *in the spirit of the special case* (16), *for every pair* $(a,b) \in \mathcal{V} \times \mathcal{V}$, $a \neq b$, *of* $\tilde{\mathbf{\Theta}}_{\text{sAvg}}$, *which is denoted by* $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}}$, *we have*

$$\frac{K^{3/2}}{\sqrt{(\sum_{k=1}^K \hat{\sigma}^2_{ab,k})(\sum_{k=1}^K 1/n_k)}}(\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}} - \mathbf{\Theta}_{ab}) = \mathbf{W}''_{ab,\mathbf{\Theta}} + \mathbf{R}''_{ab,\mathbf{\Theta}},$$

where $\mathbf{W}''_{ab,\mathbf{\Theta}}$ *converges weakly to* $\mathcal{N}(0,1)$ *and* $|\mathbf{R}''_{ab,\mathbf{\Theta}}| = o_p(1)$.

The proof is given in Appendix A.6.

Remark 5. By a similar argument to that in Remark 4, it can be shown that the asymptotic variances of $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$ and $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}}$ are both equal to σ^2_{ab}, where we recall that $\sigma^2_{ab} = \mathbf{\Theta}_{aa}\mathbf{\Theta}_{bb} + \mathbf{\Theta}^2_{ab}$. Additionally, by rewriting Theorem 3 for the full sample estimator, which uses the entire dataset of size n, it can be shown that the asymptotic variance of the debiased full estimator is also equal to σ^2_{ab}. Therefore, we can deduce that both the optimally weighted and the weighted distributed estimators are asymptotically as efficient as the debiased full sample estimator, which implies that the efficiency loss from the distributed setting is asymptotically zero.

According to the asymptotic normal distribution of the proposed estimators, one can construct confidence intervals and perform hypothesis testing for the elements of the coefficient matrix $\mathbf{\Gamma}$ and of the precision matrix $\mathbf{\Theta}$. The $(1-\alpha)100\%$ asymptotic confidence intervals for a general quantity of interest Δ, namely $\mathbf{\Theta}_{ab}$ or $\text{vec}(\mathbf{\Gamma})_a$ in this paper, using the optimally weighted, weighted and simple average estimators can be constructed as

$$\tilde{\Delta}_{\text{owAvg}} \pm \Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^K n_k/\hat{\sigma}^2_k}, \tag{23}$$

$$\tilde{\Delta}_{\text{wAvg}} \pm \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^K \hat{\sigma}^2_k)/(nK)}, \tag{24}$$

$$\tilde{\Delta}_{\text{sAvg}} \pm \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^K \hat{\sigma}^2_k)(\sum_{k=1}^K 1/n_k)/K^3}, \tag{25}$$

where $\Phi^{-1}(1-\alpha/2)$ is the $(1-\alpha/2)$-th quantile of standard normal distribution and $\hat{\sigma}_k^2$ is the k-th consistent estimator of σ^2. Substituting $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a$, $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{wAvg}})_a$ and $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{sAvg}})_a$, respectively in (23), (24) and (25) and then replacing the estimated variance $\hat{\sigma}_k^2$ by $[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}$, one can construct $(1-\alpha)100\%$ asymptotic confidence intervals for the a-th element, $a = 1,\ldots,qp$, of the vectorized from of the coefficient matrix $\mathbf{\Gamma}$. Similarly, substituting $\tilde{\mathbf{\Theta}}_{ab,\text{owAvg}}$, $\tilde{\mathbf{\Theta}}_{ab,\text{wAvg}}$ and $\tilde{\mathbf{\Theta}}_{ab,\text{sAvg}}$, respectively in (23), (24) and (25) and then replacing the estimated variance $\hat{\sigma}_k^2$ by $\hat{\sigma}_{ab,k}^2 = \hat{\mathbf{\Theta}}_{aa,k}\hat{\mathbf{\Theta}}_{bb,k} + \hat{\mathbf{\Theta}}_{ab,k}^2$, the $(1-\alpha)100\%$ asymptotic confidence intervals for every pair $(a,b) \in \mathcal{V}\times\mathcal{V}$, $a \neq b$, of the precision matrix $\mathbf{\Theta}$ can be constructed. Moreover, by substituting the same quantities in

$$|\tilde{\Delta}_{\text{owAvg}}| > \Phi^{-1}(1-\alpha/2)/\sqrt{\sum_{k=1}^{K} n_k/\hat{\sigma}_k^2},$$

$$|\tilde{\Delta}_{\text{wAvg}}| > \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}_k^2)/(nK)},$$

$$|\tilde{\Delta}_{\text{sAvg}}| > \Phi^{-1}(1-\alpha/2)\sqrt{(\sum_{k=1}^{K} \hat{\sigma}_k^2)(\sum_{k=1}^{K} 1/n_k)/K^3},$$

the rejection regions at level α for the hypothesis tests $H_0 : \text{vec}(\mathbf{\Gamma})_a = 0$ vs $H_1 : \text{vec}(\mathbf{\Gamma})_a \neq 0$, where $a = 1,\ldots,pq$ and $H_0 : \mathbf{\Theta}_{ab} = 0$ vs $H_1 : \mathbf{\Theta}_{ab} \neq 0$, where $(a,b) \in \mathcal{V}\times\mathcal{V}, a \neq b$ can be constructed.

As it is mentioned, the distributed estimator has an efficiency loss approaching zero as the sample size grows. This allows us to compare its incurred loss with that of the practically unavailable, full-sample debiased estimator. Considering the definition of $\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a$, it can be written

$$\begin{aligned}
&\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a - \text{vec}(\mathbf{\Gamma})_a \\
&= \frac{\sum_{k=1}^{K} n_k/[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K} n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} \times \sum_{k=1}^{K} \frac{\sqrt{n_k}[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{n[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\mathbf{T}_{a,k} \\
&+ \frac{\sum_{k=1}^{K} n_k/[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K} n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} \times \sum_{k=1}^{K} \frac{\sqrt{n_k}[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{n[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\mathbf{R}_{a,k,\mathbf{\Gamma}}.
\end{aligned}$$

Since $[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa} \xrightarrow{p} 1$ and $\sum_{k=1}^{K}(n_k/[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa})/$

$\sum_{k=1}^{K}(n_k/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}) \xrightarrow{p} 1$ as $K \to \infty$ and $n_k \to \infty$, $k = 1, \ldots, K$, see (66) for more details, using similar arguments to the proof of Theorem 2, we can write

$$\begin{aligned}||\tilde{\mathbf{\Gamma}}_{\text{owAvg}} - \mathbf{\Gamma}||_\infty &= \max_{a\in\{1,\ldots,qp\}} |\text{vec}(\tilde{\mathbf{\Gamma}}_{\text{owAvg}})_a - \text{vec}(\mathbf{\Gamma})_a| \\ &= O_p\big(K \max\{\sqrt{d_2 \log(pq)/n}, s_2\sqrt{\log(q)\log(pq)}/n\}\big), \quad (26)\end{aligned}$$

where the last equality is deduced using (6), (32) and (60). By using the full sample data in Theorem 2, the convergence rate of the debiased full estimator, call it $\hat{\mathbf{\Gamma}}_F^d$, is of order

$$||\hat{\mathbf{\Gamma}}_F^d - \mathbf{\Gamma}||_\infty = O_p\big(\max\{\sqrt{d_2 \log(pq)/n}, s_2\sqrt{\log(q)\log(pq)}/n\}\big). \quad (27)$$

Combining the triangle inequality with (26) and (27), it is deduced that

$$||\tilde{\mathbf{\Gamma}}_{\text{owAvg}} - \hat{\mathbf{\Gamma}}_F^d||_\infty = O_p\big(K \max\{\sqrt{d_2 \log(pq)/n}, s_2\sqrt{\log(q)\log(pq)}/n\}\big),$$

which implies that the convergence rate of the distance between the distributed estimator and the full one is equal to the rate of the distance between the distributed estimator and the true matrix $\mathbf{\Gamma}$. As such, it is noteworthy that $\tilde{\mathbf{\Gamma}}_{\text{owAvg}}$ not only follows an asymptotic normal distribution, but it also approximates the debiased full estimator $\hat{\mathbf{\Gamma}}_F^d$ well, and it exhibits a similar statistical error as $\hat{\mathbf{\Gamma}}_F^d$ as long as the number of machines K is not too large.

By a similar argument, one can derive the convergence rate of $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ to be of the order

$$\begin{aligned}||\tilde{\mathbf{\Theta}}_{\text{owAvg}} - \mathbf{\Theta}||_\infty = O_p\big(K \max\{&d_1\sqrt{\log(p)/n}, d_1^{3/2}\log(p)/n, \\ &d_1^2(\log(p))^{3/2}/(n\sqrt{n_\dagger}), d_1 s_2 \log(pq)/n\}\big),\end{aligned}$$

which is obtained by combining the definition of $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ with the rate in (12), the upper bound on $||\mathbf{W}_k||_\infty$ from Remark 3 and (36) from Appendix A.3. Comparing this convergence rate with the one from Remark 3 for the full sample data estimator, call it $\hat{\mathbf{\Theta}}_F^d$, which is of the order

$$\begin{aligned}||\hat{\mathbf{\Theta}}_F^d - \mathbf{\Theta}||_\infty = O_p\big(\max\{&d_1\sqrt{\log(p)/n}, d_1^{3/2}\log(p)/n, \\ &d_1^2(\log(p)/n)^{3/2}, d_1 s_2 \log(pq)/n\}\big),\end{aligned}$$

one can achieve the same conclusion as for the estimation of $\mathbf{\Gamma}$, which implies that if K is not too large, $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ attains a similar statistical error as the debiased full estimator $\hat{\mathbf{\Theta}}_F^d$. In the next section, the statistical error and coverage probability of estimators are compared from a finite sample perspective.

6 Simulation study

In this section, we examine empirically the performance of our proposed estimators. To this end, we followed the simulation setup of Yin and Li (2013) for generating sparse matrices $\mathbf{\Gamma}$ and $\mathbf{\Theta}$.

First, to generate the precision matrix $\mathbf{\Theta}$, we randomly generated a link between all pairs $(a, b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, with probability of connection of 0.01. Then, the corresponding entry in the precision matrix is generated uniformly from $[-1, -0.5] \bigcup [0.5, 1]$, for each link. After that, for each row, each entry except the diagonal one is divided by the sum of the absolute values of the off-diagonal entries multiplied by 1.5. Finally, the matrix is symmetrized and the diagonal entries are fixed to 1. To generate the regression coefficient matrix, we first generated a sparse indicator matrix with non-zero elements having a probability of 0.01 of occurring for every pair $(a, b); a = 1, \ldots, q, b = 1, \ldots, p$. Then, corresponding to the non-zero entries of this indicator matrix, we generated uniformly the entries of $\mathbf{\Gamma}$ from $[-1, -\nu_m] \bigcup [\nu_m, 1]$, where ν_m is the minimum absolute non-zero value of the generated precision matrix. To generate a dataset, we first generated $\dot{\mathbf{X}} = (X^1, \ldots, X^q)^\top$ from a q-dimensional Gaussian distribution with mean zero and covariance matrix $\mathbf{Q}$. We used the Toeplitz structure $\varrho^{|a-b|}$, $a, b \in \{1, \ldots, q\}$ with $\varrho = 0.9$ to generate the covariance matrix $\mathbf{Q}$. Finally, given $\dot{\mathbf{X}} = \dot{\mathbf{x}}$, we generated $\dot{\mathbf{Y}}$ from a p-dimensional Gaussian distribution with mean $\mathbf{\Gamma}^\top \dot{\mathbf{x}}$ and covariance matrix $\mathbf{\Theta}^{-1}$.

To conduct the simulation, we set $n = 25000$ and 50000 and the number of machines to $K = 5, 10$ and 20. To show the performance of the distributed estimator in the unbalanced setting, we considered the following splitting procedure. Suppose that among all available machines, two of them are powerful. The first one is the most powerful one and $(55 - K)\%$ of the dataset is distributed on this machine. The second one is less powerful than the first one and $(60 - K)\%$ of the remaining dataset is distributed on this one. The remaining dataset is distributed roughly equally on the remaining machines. By considering this splitting procedure and setting the number of responses to $p = 1100$ and the number of predictors to $q = 550$, for some sub-samples we have high-dimensionality. For example, when $n = 25000$ and $K = 10$, we have $p > n_k, \forall k = 3, \ldots, 10$ and when $K = 20$, we have $p, q > n_k$, $\forall k = 3, \ldots, 20$. Moreover, when $n = 50000$ and $K = 20$, we have $p > n_k$, $\forall k = 3, \ldots, 20$. To compare the performance of the distributed estimators, we considered two types of estimators: debiased (which are non-sparse) and

sparse. The debiased estimators consist of:

1) (Full) A debiased estimator based on the full non-distributed data.

2) (sAvg) An estimator based on first splitting the data and averaging directly the debiased estimators from each machine.

3) (wAvg) An estimator based on splitting first the data and taking the weighted average of the debiased estimators from each machine, where the weight for the k-th sub-sample is set to (n_k/n).

4) (Top1) The estimator produced by the most powerful machine which takes $(55-K)\%$ of dataset. Since estimation on each machine is consistent and asymptotically normal, investigating the performance on the first machine which takes most of the dataset, is relevant.

The sparse competitors, which are shown respectively by SFull, SsAvg, SwAvg, STop1, are obtained in the same way as estimators in 1)-4) but without the debiasing step. A comparison with the full estimator reveals how much the performance deteriorates due to splitting the data, while a comparison with the simple and weighted average estimators has the purpose of evaluating if indeed the proposed estimator is better equipped to tackle unbalanced settings due to a more appropriate weighting. A comparison with the Top1 estimator has the purpose to evaluate if the remaining $(K-1)$ machines which account for $(45+K)\%$ of the original data are still able to produce informative estimates even though they receive low amounts of data. In this study, the tuning parameters in (3), (8) and (28), which is needed to obtain $\mathbf{M}_k$, see Appendix A.1, are set to $\rho_k = \rho = \sqrt{\log(pq)/n}$, $\lambda_k = \lambda = \sqrt{\log(p)/n}$ and $\tilde{\rho}_{j,k} = \tilde{\rho} = \sqrt{\log(q)/n}$ for all simulation runs. All simulation results are calculated as averages over $R = 500$ different replications.

To compare the performance of the estimators, we used the Frobenius norm between the estimated matrix for each competitor and the true matrix from the data generating process. The results of the Frobenius norm and their standard deviation (between parentheses) for $n = 50000$ are presented in Table 1 for each competitor, separately on the active and the non-active sets. Note that the active sets on $\mathbf{\Theta}$ and $\mathbf{\Gamma}$ are indexed by S_1 and S_2, respectively. The results for $n = 25000$ are similar and are presented in Table 4 in Appendix C.

From Table 1, it is observed that the performance of the proposed debiased owAvg estimator is similar to that of the non-distributed, full estimator

in terms of Frobenius norm. With increasing K from 5 to 20, the norm of the distributed estimator stays relatively constant. This suggests that by splitting observations in combination with the proposed aggregation, one might not lose much information. On the other hand, wAvg is quite sensitive to the number of machines and with increasing K, its norm performance deteriorates. Especially, when $K = 20$, the difference relative to the full one is large on the non-active set. The worst estimator between debiased ones is sAvg which imposes the same weights on all sub-samples and it is observed that its norm increases substantially, which suggests that it is highly sensitive to K, as opposed to the distributed owAvg estimator. Furthermore, the Frobenius norm of Top1 is much larger than that of the centralized full estimator, which implies that by considering just the first machine with the largest amount of data and disregarding the remaining machines one loses information as this strategy does not provide an accurate estimate. Moreover, as it is expected, the performance of the sparse estimators is much better than the performance of the debiased estimators on the non-active set as they shrink most of the elements to zero. However, their errors are much larger than the errors of the debiased estimators on the active set as they do not correct for the bias.

Due to the asymptotic distribution of the proposed estimators, investigating their inferential properties is also of interest. However, since there is no distributional result available for the sparse estimators, it is not possible to perform inference using these estimators. Figure 1 shows the normalized distribution of the proposed debiased optimally weighted estimator and the full non-distributed estimator for both $\mathbf{\Gamma}$ and $\mathbf{\Theta}$, respectively from top to bottom. As an illustrative example, these figures are reported for $(a, b) = (1, 3)$, others are available from the authors, but are similar. The light gray histograms correspond to the normalized distribution of the full estimators which are obtained from (6) and (13), by setting $K = 1$, for $\mathbf{\Gamma}$ and $\mathbf{\Theta}$, respectively. The dark gray histograms correspond to the asymptotic distributions of (21) and (22), respectively. The presented histograms confirm the asymptotic normal distribution of the proposed estimators and its similarity to the full one.

Using the asymptotic distribution of the estimators, the coverage probabilities and the length of the confidence intervals are presented in Table 2 at significance level $\alpha = .05$. Here, we explain the procedure for computing the empirical coverage probability for the elements of $\mathbf{\Theta}$. The same procedure is applied to compute the empirical coverage probability and the length of the confidence intervals for the elements of $\mathbf{\Gamma}$. Similarly to Jankova and

Table 1: Average and standard deviation (between parentheses) of the Frobenius norm over 500 repetitions on the active and non-active sets, for the proposed estimators and different competitors, when $n = 50000$.

			Active set				Non-active set			
					K				K	
			1	5	10	20	1	5	10	20
$\boldsymbol{\Theta}$	Debiased	Full	0.73				4.75			
			(.01)				(.01)			
		owAvg		0.74	0.78	0.89		4.86	5.13	5.59
				(.00)	(.01)	(.01)		(.00)	(.01)	(.01)
		Top1		1.00	1.05	1.19		6.60	6.95	7.85
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		sAvg		1.07	1.81	2.94		7.00	10.44	12.96
				(.01)	(.02)	(.02)		(.01)	(.05)	(.03)
		wAvg		0.75	0.84	1.32		4.89	5.42	6.74
				(.00)	(.01)	(.01)		(.00)	(.01)	(.01)
	Sparse	SFull	2.03				.38			
			(.01)				(.00)			
		STop1		2.13	2.15	2.20		1.02	1.19	1.69
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		SsAvg		2.02	1.82	1.49		3.12	5.06	6.06
				(.01)	(.01)	(.01)		(.00)	(.02)	(.01)
		SwAvg		1.99	1.90	1.68		1.44	1.95	2.78
				(.01)	(.01)	(.01)		(.00)	(.00)	(.00)
$\boldsymbol{\Gamma}$	Debiased	Full	1.09				10.85			
			(.01)				(.01)			
		owAvg		1.11	1.15	1.23		11.10	11.44	12.24
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		Top1		1.55	1.63	1.86		15.43	16.29	18.54
				(.01)	(.01)	(.02)		(.02)	(.02)	(.02)
		sAvg		1.57	2.02	2.23		15.67	20.15	22.23
				(.01)	(.02)	(.02)		(.02)	(.03)	(.03)
		wAvg		1.11	1.16	1.29		11.12	11.59	12.89
				(.01)	(.01)	(.01)		(.01)	(.01)	(.02)
	Sparse	SFull	4.74				1.93			
			(.00)				(.00)			
		STop1		4.74	4.74	4.75		1.99	2.00	2.04
				(.00)	(.00)	(.00)		(.00)	(.00)	(.00)
		SsAvg		4.75	4.36	3.97		2.04	2.27	2.65
				(.00)	(.13)	(.03)		(.00)	(.05)	(.02)
		SwAvg		4.74	4.61	4.40		1.93	1.90	1.93
				(.00)	(.04)	(.01)		(.00)	(.01)	(.00)

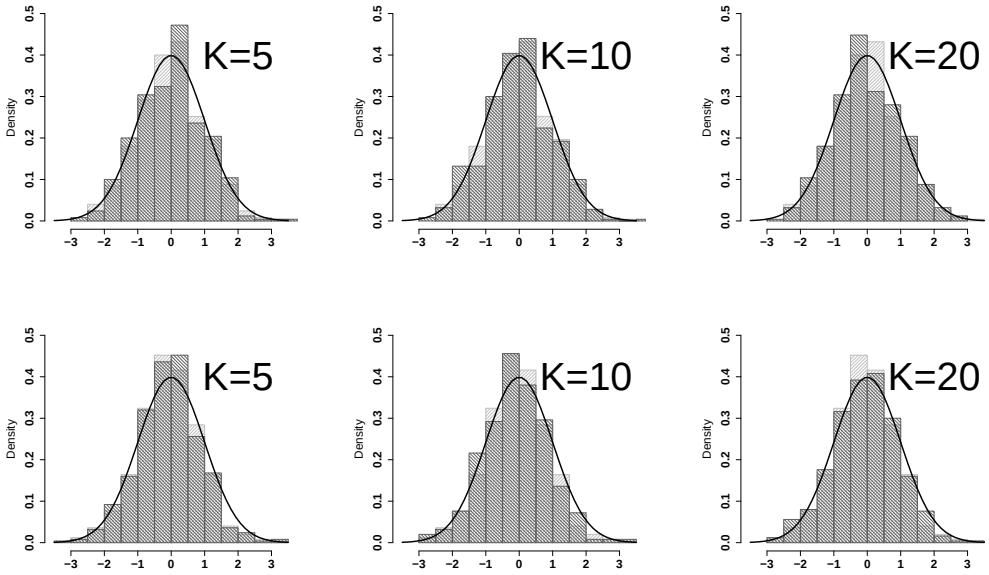

Figure 1: Histogram of the normalized full (light gray bars) and distributed (dark gray bars), respectively, when $n = 50000$, $(a, b) = (1, 3)$ and $K = 5,\ 10$ or 20. From top to bottom, the figures present the asymptotic distributions for the estimation of $\mathbf{\Gamma}$ (top) and $\mathbf{\Theta}$ (bottom).

van de Geer (2015), the empirical probability that the true parameter $\mathbf{\Theta}_{ab}$ is included in the confidence interval is defined as $\hat{\mathbb{P}}_{ab} = \#\{\mathbf{\Theta}_{ab} \in \mathrm{CI}_{ab,r}\}/R$, where R is the number of replications in the simulation, $\mathrm{CI}_{ab,r}$ is the estimated confidence interval for $\mathbf{\Theta}_{ab}$ at the r-th replication, and # denotes the number of times in which the true parameter $\mathbf{\Theta}_{ab}$ belongs to the confidence interval. After obtaining $\hat{\mathbb{P}}_{ab}$ for all $(a, b) \in \mathcal{V} \times \mathcal{V}, a \neq b$, the average coverage probability on the active set S_1 is obtained as $\mathrm{Avg.Cov}_{S_1} = (1/s_1) \sum_{(a,b)\in S_1} \hat{\mathbb{P}}_{ab}$, where s_1 is the number of active components. Similar computations have been implemented for obtaining the estimated coverage probability over the non-active set S_1^c.

Table 2: Average coverage probability and average length of the confidence intervals over 500 repetitions for the proposed estimators and different competitors, when $n = 50000$.

			Avg.Cov				Avg.Len			
			K				K			
			1	5	10	20	1	5	10	20
$\boldsymbol{\Theta}$	Active set	Full	.93				.02			
		owAvg		.93	.93	.92		.02	.02	.02
		Top1		.94	.94	.94		.02	.03	.02
		sAvg		.92	.85	.71		.02	.03	.04
		wAvg		.94	.96	.91		.02	.02	.03
	Non-active set	Full	.93				.02			
		owAvg		.95	.95	.95		.02	.02	.02
		Top1		.95	.95	.95		.02	.03	.05
		sAvg		.94	.92	.92		.02	.03	.04
		wAvg		.95	.97	.98		.02	.02	.03
$\boldsymbol{\Gamma}$	Active set	Full	.95				.06			
		owAvg		.95	.95	.95		.06	.06	.06
		Top1		.95	.95	.95		.09	.08	.09
		sAvg		.95	.94	.93		.09	.10	.10
		wAvg		.96	.96	.97		.06	.06	.07
	Non-active set	Full	.96				.08			
		owAvg		.95	.95	.95		.06	.06	.06
		Top1		.95	.95	.95		.09	.08	.09
		sAvg		.95	.94	.93		.09	.10	.10
		wAvg		.96	.96	.97		.06	.06	.07

From Table 2, it is observed that the performance of the owAvg estimator is close to the full one on both the active and non-active sets, as the coverage probabilities are close to the nominal level of 95%. Moreover, the average lengths are relatively low and are stable with increasing K. The results for Top1 are also close to the nominal level, but its length is slightly larger than that of the full and of the distributed estimator. However, in Table 1, we observed that its Frobenius norm is far away from the norm of the full estimator making it a less interesting alternative. The coverage probability

of sAvg is low in some cases, especially in estimating $\mathbf{\Theta}$ on the active set. The length of its confidence interval is also relatively large, especially when estimating $\mathbf{\Gamma}$. The performance of wAvg is generally better than that of sAvg but over-coverage is observed. More than that, the length of its confidence interval is not stable and it increases with increasing K. The results for $n = 25000$ are similar and they are presented in Table 5 in Appendix C.

Another quantity which is important to keep track of, is the running time. In this paper, it is considered as the maximum running time among all parallel jobs plus the time to combine the results. These results are shown in Figure 2 for the debiased estimators of $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ for different sample sizes from $n = 50000$ to 200000 and $K = 5, 10, 20$. Not surprisingly, the running time of the sparse estimators were less than the running time of the debiased ones as they do not have the debiasing step in the estimation procedure, and they are not reported in this figure. It is observed from Figure 2 that for any fixed sample size, the computation time of the distributed estimators is less than that of the full one and as expected it decreases with increasing K. This running time is quite close for all owAvg, wAvg, sAvg and Top1 estimators. This behavior is the same for both estimators of $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ and shows, as expected, the efficiency of the proposed estimators in terms of computation time.

7 Real data example

To explore the performance of the proposed methodology, we used the Pan-Cancer dataset from The Cancer Genome Atlas (TCGA) project (available at https://xenabrowser.net/datapages/). This project was started in 2006 and in 10 years time, TCGA network investigators had characterized the molecular landscape of tumors from more than 11000 patients across 33 cancer types. This particular TCGA molecular dataset has been studied in multiple works to understand the cancer biology, including those of glioblastoma multiforme (GBM), ovarian, breast, lung, prostate, bladder and others (see for instance, Akbani et al., 2015; Cancer Genome Atlas Research Network, 2017; Liu et al., 2018).

Although the estimation of the graph structure of genes can be effective in identifying the associated genetic variants, external covariates such as single nucleotide polymorphisms may affect their structure. As such, we applied the proposed conditional multivariate regression to regress 743 microRNA

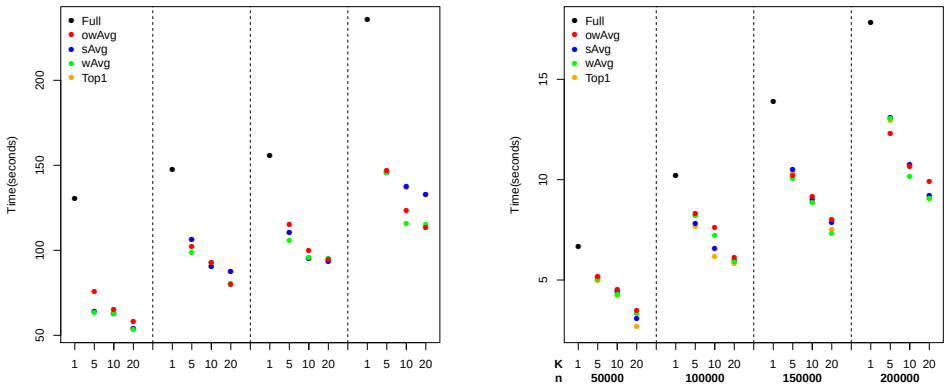

Figure 2: Running time in seconds for the full and proposed estimators in estimating $\mathbf{\Gamma}$ (left) and $\mathbf{\Theta}$ (right), when $p = 1100$ and $q = 550$. The regularization parameters for the distributed estimators are considered as $\lambda_k = \sqrt{\log(p)/n_k}$, $\rho_k = \sqrt{\log(pq)/n_k}$ and $\tilde{\rho}_k = \sqrt{\log(q)/n_k}$.

mature strand expressions on 27147 tumor gene-level copy numbers and to model how the microRNA expressions regulate gene expressions. Estimation of the underlying graph among the microRNAs is also of interest. We further selected 1164 tumor gene expressions with variance greater than 0.3, and the final dataset consists of $n = 9986$ subjects having both $q = 1164$ covariates and $p = 743$ responses.

To compare the performance of the proposed method, we split the dataset on $K = 3, 5$ and 10 different machines with the same splitting setting as in the simulation study in Section 6. Tuning parameters are also fixed as in the simulation study with $\alpha = 5\%$ and a Bonferroni correction is applied for multiple testing. The results are presented in Table 3 and it is observed that when estimating $\mathbf{\Gamma}$, the owAvg and wAvg estimators identify more common non-zero coefficients with the full estimator, while sAvg and Top1 are further away from the full estimator. When $K = 3$, sAvg is close to the full one, but with increasing K, it grows further apart, such that when $K = 10$, it identified only 234 coefficients of which only 217 of them are common with the full estimator. The same holds for Top1 which identified much less coefficients compared to the full, owAvg and wAvg estimators. We conclude that there

Table 3: Significant and common pairs between four competitors and the estimator using the full dataset. For all methods a Bonferroni correction is applied.

		Significant pairs				Common pairs with Full		
		K				K		
		1	3	5	10	3	5	10
$\mathbf{\Gamma}$	Full	964						
	owAvg		897	832	828	797	756	685
	Top1		300	290	295	255	241	186
	sAvg		817	466	234	735	416	217
	wAvg		888	948	1334	812	761	778
$\mathbf{\Theta}$	Full	6564						
	owAvg		6168	5917	5535	5888	5675	5299
	Top1		3095	2966	2655	3060	2933	2632
	sAvg		5234	2923	1914	5099	2896	1902
	wAvg		6120	5463	4169	5854	5331	4130

are more common edges between the graphs estimated by the distributed owAvg method and the full one, while the least similar competitors are Top1 and sAvg. With increasing K, the number of common edges tends to decrease for all competitors due to the loss of information by splitting data on more machines.

As it is mentioned in Wang (2015), the tuple hsa-miR-136, hsa-miR-376a and hsa-miR-377 are the most important genes in identifying GBM cancer. Top1 could not identify any edges between these genes, while sAvg could identify an edge between hsa-miR-136 and hsa-miR-377, when $K = 3$. With the full, distributed owAvg and wAvg estimators, we could successfully identify an edge between hsa-miR-136 and hsa-miR-377 and another one between hsa-miR-376a and hsa-miR-377. However, wAvg missed the couple hsa-miR-376a and hsa-miR-377 when $K = 10$. The sparse graphical Lasso estimator is also considered as a competitor and with this method we identified 37376 edges suggesting that many estimated edges might in fact be false positive edges.

Afterwards, to investigate the performance of the estimators for completely independent variables, we permuted randomly sample data for each variable, thus breaking up the correlation structure in order to construct a dataset with mutually independent variables. As such, we expect that there are no selected variables in the estimated coefficient matrix and zero

off-diagonal elements in the estimated precision matrix. By implementing separately the proposed owAvg estimator and all competitors on the dataset with $K = 3, 5$ and 10, we identified no edges in the estimated precision matrix which confirms this assertion. However, a couple of non-zero coefficients were wrongly identified in the estimated coefficient matrix. The owAvg and sAvg estimators, both identified around 10 non-zero coefficients, while wAvg identified 84, which indicates more false positive discoveries produced by this estimator.

8 Discussion

Splitting a dataset on multiple locations with different sizes is an inevitable method to tackle the problem of large scale datasets which cannot be read nor stored in one single location. It is also of contemporary interest due to today's security and privacy concerns. The most essential step in distributed problems is choosing how to aggregate different estimators to a final one.

In this paper, aggregated estimators are introduced for the coefficient matrix in the multivariate regression models and the precision matrix corresponding to the graph structure of the response vector. To build these estimators, first debiased estimators were provided on each machine and then, they were pooled together to create the final estimators by maximizing a pseudo log-likelihood function which is constructed using the asymptotic distribution of the debiased estimators. Two special cases of the aggregated estimators, including simple and weighted averages, were provided under the known variance assumption. Statistical guarantees and the asymptotic distribution of the aggregated estimators were investigated under sparsity conditions and a growing number of machines as a function of the sample size.

It is shown that the aggregated estimators are asymptotically as efficient as the debiased, full non-distributed estimators and confirm the asymptotic negligibility of the efficiency loss in the distributed procedure. Moreover, based on the provided convergence rates, it is deduced that the aggregated estimators exhibit a similar statistical error as the debiased full estimator as long as the number of machine is not too large. In the estimation of the coefficient matrix, it is shown that as the number of machines grows at the rate of $K = O\big(n^{\pi_4}/(\sqrt{\log(pq)\log(q)}\max\{s_2, \sqrt{d_2}\})\big)$, $0 < \pi_4 < 1/4$, the final estimator is consistent and asymptotically normal. This growth rate

adjusts to $K = O\big(n^{\pi_7}/(d_1 \log(p))\big)$, $0 < \pi_7 \leq 1/3$, in estimating the precision matrix. Statistical inference was also proposed based on the asymptotic normal distribution of the aggregated estimators. These distributions are valid as long as the number of machines grows at the mentioned rates.

The finite sample performance of the proposed estimators was evaluated with a simulation study where it was observed that the estimators produced competitive results relative to the non-distributed estimators that use the entire data. Since we perform estimation by distributing the computational load across multiple machines, not surprisingly the computational time comparison favors our novel estimator. Moreover, this estimator performed substantially better than the simple average-based estimator in terms of accuracy. It was also observed that the coverage probabilities of the distributed estimators are close to those of the non-distributed estimators. This points to the fact that in practice, performing distributed estimation across multiple machines in our unbalanced framework induces a minimal loss in performance relative to models using all the data in a centralized location.

Acknowledgement

Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

A Proofs of the main theorems and lemmas

A.1 Proof of Theorem 1

Before starting the proof of Theorem 1, we provide a short description of the nodewise Lasso method from van de Geer et al. (2014) using the k-th sub-sample, which is needed later in the proof.

For every $j \in \{1, \ldots, q\}$, use Lasso for the regression problem $\mathbf{X}_{j,k}$ against $\mathbf{X}_{-j,k}$, where $\mathbf{X}_{j,k}$ and $\mathbf{X}_{-j,k}$ are, the j-th column and $n_k \times (q-1)$ dimensional sub-matrix of $\mathbf{X}_k$ obtained by removing the j-th column, respectively. Formally,

$$\hat{\eta}_{j,k} = \arg \min_{\eta \in \mathbb{R}^{q-1}} \left\{ \frac{1}{2n_k} ||\mathbf{X}_{j,k} - \mathbf{X}_{-j,k}\eta||_2^2 + \tilde{\rho}_{j,k}||\eta||_1 \right\}, \tag{28}$$

with components $\hat{\eta}_{j,k} = \{\hat{\eta}_{(j,j'),k}; j' = 1, \ldots, q, j' \neq j\}$. Define

$$\hat{\mathbf{\Psi}}_k = \begin{pmatrix} 1 & -\hat{\eta}_{(1,2),k} & \ldots & -\hat{\eta}_{(1,q),k} \\ -\hat{\eta}_{(2,1),k} & 1 & \ldots & -\hat{\eta}_{(2,q),k} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\eta}_{(q,1),k} & -\hat{\eta}_{(q,2),k} & \ldots & 1 \end{pmatrix}$$

and write

$$\hat{\mathbf{\Upsilon}}_k^2 = \text{diag}(\hat{\tau}_{1,k}^2, \ldots, \hat{\tau}_{q,k}^2), \quad \hat{\tau}_{j,k}^2 = ||\mathbf{X}_{j,k} - \mathbf{X}_{-j,k}\hat{\eta}_{j,k}||_2^2/n_k + \tilde{\rho}_{j,k}||\hat{\eta}_{j,k}||_1,$$

and finally, set

$$\mathbf{M}_k = \hat{\mathbf{\Upsilon}}_k^{-2}\hat{\mathbf{\Psi}}_k.$$

Now to prove Gaussianity in Theorem 1, we mention that ξ_k is an $n_k \times p$ matrix with independent rows and distributed as $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$. As such, the random vector $\text{vec}(\xi_k)$ is distributed as a pn_k-dimensional Gaussian vector with mean zero and covariance matrix $(\mathbf{\Sigma} \otimes \mathbf{I}_{n_k})$, which is a $pn_k \times pn_k$ matrix. Due to the properties of the $\text{vec}(\cdot)$ function,

$$\text{vec}(\mathbf{M}_k\mathbf{X}_k^\top\xi_k/\sqrt{n_k}) = (1/\sqrt{n_k})\big(\mathbf{I}_p \otimes (\mathbf{M}_k\mathbf{X}_k^\top)\big)\text{vec}(\xi_k).$$

As such, given $\mathbf{X}_k$, the random vector $\mathbf{T}_k$ is a linear transformation of $\text{vec}(\xi_k)$ and thus it is a pq-dimensional Gaussian vector with mean zero and variance

$$\begin{aligned} \mathbb{V}\text{ar}(\mathbf{T}_k|\mathbf{X}_k) &= (1/n_k)(\mathbf{I}_p \otimes (\mathbf{M}_k\mathbf{X}_k^\top))(\mathbf{\Sigma} \otimes \mathbf{I}_{n_k})(\mathbf{I}_p \otimes (\mathbf{X}_k\mathbf{M}_k^\top)) \\ &= \mathbf{\Sigma} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top). \end{aligned}$$

To show negligibility of the remainder term $\mathbf{R}_{k,\mathbf{\Gamma}}$ in (5), we have

$$\begin{aligned} ||\mathbf{R}_{k,\mathbf{\Gamma}}||_\infty &\leq \sqrt{n_k}||\mathbf{I}_q - \mathbf{M}_k\mathbf{C}_k||_\infty|||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}|||_\infty \\ &\leq \sqrt{n_k}||\mathbf{I}_q - \mathbf{M}_k\mathbf{C}_k||_\infty||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}||_1. \end{aligned} \tag{29}$$

Using the KKT conditions, van de Geer et al. (2014) showed that $||\mathbf{C}_k\mathbf{M}_{j,k}^\top - \mathbf{e}_j||_\infty \leq \tilde{\rho}_{j,k}/\hat{\tau}_{j,k}^2$, where $\mathbf{M}_{j,k}$ is the j-th row of $\mathbf{M}_k$ and $\mathbf{e}_j$ is the j-th unit column vector with 1 at the j-th position and zero at others. Under the assumptions (A1) and (A2) and considering the maximum row sparsity $d_2 = o(\sqrt{n_\dagger}/\log(q))$ in Lemma 5.3 of van de Geer et al. (2014), we have

$$\max_{j \in \{1,\ldots,q\}} 1/\hat{\tau}_{j,k}^2 = O_p(1).$$

Therefore, by choosing uniformly $\tilde{\rho}_{j,k} \asymp \sqrt{\log(q)/n_k}$ for each $j = 1, \ldots, q$, we get

$$||\mathbf{C}_k\mathbf{M}_k^\top - \mathbf{I}_q||_\infty = \max_j ||\mathbf{C}_k\mathbf{M}_{j,k} - \mathbf{e}_j||_\infty = O_p(\sqrt{\log(q)/n_k}). \tag{30}$$

Substituting (30) and (54) from Lemma 4 (see Appendix B) in (29), we get

$$||\mathbf{R}_{k,\boldsymbol{\Gamma}}||_\infty = O_p\big(s_2\sqrt{\log(q)\log(pq)/n_k}\big).$$

Finally, by considering $s_2 = o\big(n_\dagger^{\pi_1}/(\log(q)\log(pq))\big)$, $0 < \pi_1 \leq 1/2$, the required result follows. □

A.2 Proof of Theorem 2

Using (5), we have that,

$$||\hat{\boldsymbol{\Gamma}}_k^d - \boldsymbol{\Gamma}||_\infty \leq ||\text{vec}(\mathbf{M}_k\mathbf{X}_k^\top\xi_k)||_\infty/n_k + ||\mathbf{R}_{k,\boldsymbol{\Gamma}}||_\infty/\sqrt{n_k}. \tag{31}$$

Due to the properties of the vec$(\cdot)$ function,

$$\begin{aligned} ||\text{vec}(\mathbf{M}_k\mathbf{X}_k^\top\xi_k)||_\infty/n_k &\leq |||\mathbf{I}_p \otimes \mathbf{M}_k|||_\infty||\text{vec}(\mathbf{X}_k^\top\xi_k)||_\infty/n_k \\ &= O_p(\sqrt{d_2\log(pq)/n_k}), \end{aligned} \tag{32}$$

where the last equality is obtained by (i) working on the event $\mathcal{F}_k(n_k, p, q)$ with $\rho_k \asymp \sqrt{\log(pq)/n_k}$, and (ii) by the same argument as in the proof of Lemma 5.4 from van de Geer et al. (2014), as

$$|||\mathbf{I}_p \otimes \mathbf{M}_k|||_\infty = |||\mathbf{M}_k|||_\infty = \max_{j\in\{1,\ldots,q\}} ||\mathbf{M}_{j,k}||_1 = O_p(\sqrt{d_2}),$$

where $\mathbf{M}_{j,k}$ is the j-th row of $\mathbf{M}_k$. Under assumptions (A1) and (A2) and using Theorem 1, by substituting (6) and (32) in (31), the result in (7) follows directly. □

A.3 Proof of Lemma 1

Before starting the proof of this Lemma, we provide some technical assumptions on the sample covariance matrix $\mathbf{C}_k$, which are needed later in the proof. For more information on these assumptions, the reader can refer to Yin and Li (2013).

(C1) There exists an $\alpha_2 \in (0, 1]$, such that

$$\sup_{b\in\{1,\ldots,p\}} |||(\mathbf{C}_k)_{S_2^c(b)S_2(b)}[(\mathbf{C}_k)_{S_2(b)S_2(b)}]^{-1}|||_\infty \leq 1 - \alpha_2,$$

where $(\mathbf{C}_k)_{S_2^c(b)S_2(b)}$ is a sub-matrix of $\mathbf{C}_k$ whose rows and columns are indexed by the elements of $S_2^c(b)$ and $S_2(b)$, respectively, where $S_2^c(b)$ is the complement set of $S_2(b)$, defined in Section 2.

(C2) There exists a constant $C_{\max}$, such that the largest eigenvalue

$$\Lambda_{\max}\bigg(\big[(\mathbf{C}_k \otimes \mathbf{I}_p)_{S_2S_2}\big]^{-1}(\mathbf{C}_k \otimes \mathbf{\Sigma})_{S_2S_2}\big[(\mathbf{C}_k \otimes \mathbf{I}_p)_{S_2S_2}\big]^{-1}\bigg) \leq C_{\max}.$$

(C3) For all $n_k > 0$, the largest eigenvalue of $\mathbf{C}_k$ has a common upper bound Λ_3, that is $\Lambda_{\max}(\mathbf{C}_k) \leq \Lambda_3$.

Now to prove Lemma 1, using (11), we have

$$\begin{aligned}
||\mathbf{R}_{k,\mathbf{\Theta}}||_\infty &= \sqrt{n_k}||-(\hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{I}_p)(\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}) - (\hat{\mathbf{\Theta}}_k - \mathbf{\Theta})\mathbf{W}_k^{'}\mathbf{\Theta} - \mathbf{\Theta}\mathbf{W}_k^{''}\mathbf{\Theta}||_\infty \\
&\leq \sqrt{n_k}||\hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{I}_p||_\infty|||\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}|||_\infty + \sqrt{n_k}|||\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}|||_\infty||\mathbf{W}_k^{'}\mathbf{\Theta}||_\infty \\
&\quad + \sqrt{n_k}||\mathbf{\Theta}\mathbf{W}_k^{''}\mathbf{\Theta}||_\infty.
\end{aligned} \tag{33}$$

To find an upper bound on $||\hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{I}_p||_\infty$, we write

$$\begin{aligned}
||\hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{I}_p||_\infty &= ||\hat{\mathbf{\Theta}}_k\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \hat{\mathbf{\Theta}}_k\mathbf{\Sigma} + \hat{\mathbf{\Theta}}_k\mathbf{\Sigma} - \mathbf{I}_p||_\infty \\
&= ||\hat{\mathbf{\Theta}}_k(\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}) + \hat{\mathbf{\Theta}}_k\mathbf{\Sigma} - \mathbf{\Theta}\mathbf{\Sigma} + \mathbf{\Theta}\mathbf{\Sigma} - \mathbf{I}_p||_\infty \\
&= ||\hat{\mathbf{\Theta}}_k(\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}) + (\hat{\mathbf{\Theta}}_k - \mathbf{\Theta})\mathbf{\Sigma} + \mathbf{\Theta}\mathbf{\Sigma} \\
&\quad - \mathbf{\Theta}\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} + \mathbf{\Theta}\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{I}_p||_\infty \\
&= ||\hat{\mathbf{\Theta}}_k(\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}) + (\hat{\mathbf{\Theta}}_k - \mathbf{\Theta})\mathbf{\Sigma} - \mathbf{\Theta}(\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}) \\
&\quad + \mathbf{\Theta}\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Theta}\mathbf{\Sigma}||_\infty \\
&\leq |||\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}|||_\infty||\mathbf{W}_k^{'}||_\infty + ||\hat{\mathbf{\Theta}}_k - \mathbf{\Theta}||_\infty|||\mathbf{\Sigma}|||_\infty + ||\mathbf{\Theta}\mathbf{W}_k^{'}||_\infty.
\end{aligned} \tag{34}$$

Substituting (34) in (33), we have

$$||\mathbf{R}_{k,\mathbf{\Theta}}||_\infty \leq \sqrt{n_k}\big\{2||\mathbf{R}_{k,\mathbf{\Theta}}^1||_\infty + ||\mathbf{R}_{k,\mathbf{\Theta}}^2||_\infty + ||\mathbf{R}_{k,\mathbf{\Theta}}^3||_\infty + ||\mathbf{R}_{k,\mathbf{\Theta}}^4||_\infty\big\}, \tag{35}$$

where

$$
\begin{aligned}
||\mathbf{R}^1_{k,\boldsymbol{\Theta}}||_\infty &= ||\boldsymbol{\Theta}\mathbf{W}_k^{'}||_\infty |||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}|||_\infty, \\
||\mathbf{R}^2_{k,\boldsymbol{\Theta}}||_\infty &= |||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}|||^2_\infty ||\mathbf{W}_k^{'}||_\infty, \\
||\mathbf{R}^3_{k,\boldsymbol{\Theta}}||_\infty &= ||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}||_\infty |||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}|||_\infty \kappa_{\boldsymbol{\Sigma}}, \\
||\mathbf{R}^4_{k,\boldsymbol{\Theta}}||_\infty &= ||\boldsymbol{\Theta}\mathbf{W}_k^{''}\boldsymbol{\Theta}||_\infty.
\end{aligned}
$$

Under assumption (B1) for the matrix $\boldsymbol{\Theta}$, it holds that

$$
|||\boldsymbol{\Theta}|||_\infty = \max_{a\in\mathcal{V}} ||\boldsymbol{\Theta}_a||_1 \leq \sqrt{d_1} \max_{a\in\mathcal{V}} ||\boldsymbol{\Theta}_a||_2 \leq \sqrt{d_1}\Lambda_{\max}(\boldsymbol{\Theta}) = O(\sqrt{d_1}), \quad (36)
$$

where $\boldsymbol{\Theta}_a$ is the a-th row of $\boldsymbol{\Theta}$. Under (B2) and additional assumptions (C1)-(C3), the conditions of result 1 from Theorem 2 of Yin and Li (2013) are fulfilled, and for the k-th sub-sample we have

$$
||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}||_\infty = O_p\left(\left\{ 16\sqrt{2}(1+4\gamma^2)(1+8/\alpha_1) \max_a \boldsymbol{\Sigma}_{aa}\kappa_{\mathbf{H}} \right\} \sqrt{\frac{\log(4p^\tau)}{n_k}} \right), \quad (37)
$$

where $\max_a \boldsymbol{\Sigma}_{aa}$ is the maximal diagonal element of the covariance matrix $\boldsymbol{\Sigma}$, $\tau > 2$ is a constant and γ is a common sub-Gaussian parameter for Gaussian random variables $\varepsilon^1, \cdots, \varepsilon^p$, where $\gamma = O(1)$. Moreover, based on result 2 of the aforementioned theorem, the edge set $\mathcal{E}(\hat{\boldsymbol{\Theta}}_k)$, i.e. the edge set created based on the estimated $\hat{\boldsymbol{\Theta}}_k$, is a subset of the true edge set $\mathcal{E}(\boldsymbol{\Theta})$ with high probability. As such, $\hat{\boldsymbol{\Theta}}_k$ has at most d_1 non-zero entries per row, and we get

$$
\begin{aligned}
|||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}|||_\infty = \max_a \sum_{b=1}^{p} |\hat{\boldsymbol{\Theta}}_{ab,k} - \boldsymbol{\Theta}_{ab}| &= \max_a \sum_{b=1}^{d_1} |\hat{\boldsymbol{\Theta}}_{ab,k} - \boldsymbol{\Theta}_{ab}| \\
&\leq \max_a \sum_{b=1}^{d_1} \max_b |\hat{\boldsymbol{\Theta}}_{ab,k} - \boldsymbol{\Theta}_{ab}| \\
&= d_1 ||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}||_\infty,
\end{aligned}
$$

where $\hat{\boldsymbol{\Theta}}_{ab,k}$ is the (a,b)-th element of $\hat{\boldsymbol{\Theta}}_k$. Thus, using (37),

$$
|||\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}|||_\infty \leq O_p\left(d_1 \left\{ 16\sqrt{2}(1+4\gamma^2)(1+8/\alpha_1) \max_a \boldsymbol{\Sigma}_{aa}\kappa_{\mathbf{H}} \right\} \sqrt{\frac{\log(4p^\tau)}{n_k}} \right).
$$

Therefore in (35), we have

$$||\mathbf{R}_{k,\boldsymbol{\Theta}}||_\infty \leq 5\sqrt{n_k}\max\big\{||\mathbf{R}^1_{k,\boldsymbol{\Theta}}||_\infty, ||\mathbf{R}^2_{k,\boldsymbol{\Theta}}||_\infty, ||\mathbf{R}^3_{k,\boldsymbol{\Theta}}||_\infty, ||\mathbf{R}^4_{k,\boldsymbol{\Theta}}||_\infty\big\}, \quad (38)$$

where

$$\begin{aligned}
||\mathbf{R}^1_{k,\boldsymbol{\Theta}}||_\infty &\leq \sqrt{d_1}\Lambda_{\max}(\boldsymbol{\Theta})||\mathbf{W}^{'}_k||_\infty \\
&\quad \times O_p\bigg(d_1\big\{16\sqrt{2}(1+4\gamma^2)(1+8/\alpha_1)\max_a \boldsymbol{\Sigma}_{aa}\kappa_{\mathbf{H}}\big\}\sqrt{\frac{\log(4p^\tau)}{n_k}}\bigg), \\
||\mathbf{R}^2_{k,\boldsymbol{\Theta}}||_\infty &\leq ||\mathbf{W}^{'}_k||_\infty O_p\bigg(d_1^2\big\{16\sqrt{2}(1+4\gamma^2)(1+8/\alpha_1)\max_a \boldsymbol{\Sigma}_{aa}\kappa_{\mathbf{H}}\big\}^2\frac{\log(4p^\tau)}{n_k}\bigg), \\
||\mathbf{R}^3_{k,\boldsymbol{\Theta}}||_\infty &\leq \kappa_{\boldsymbol{\Sigma}} O_p\bigg(d_1\big\{16\sqrt{2}(1+4\gamma^2)(1+8/\alpha_1)\max_a \boldsymbol{\Sigma}_{aa}\kappa_{\mathbf{H}}\big\}^2\frac{\log(4p^\tau)}{n_k}\bigg).
\end{aligned}$$

To obtain an upper bound on $||\mathbf{W}^{'}_k||_\infty$, using Lemma 2 of Yin and Li (2013), we can write

$$||\mathbf{W}^{'}_k||_\infty = O_p\bigg(\sqrt{\frac{\log(4p^\tau)}{C_2 n_k}}\bigg),$$

where $C_2 = [128(1+4\gamma^2)^2\max_a^2 \boldsymbol{\Sigma}_{aa}]^{-1}$. Moreover,

$$||\mathbf{R}^4_{k,\boldsymbol{\Theta}}||_\infty = ||\boldsymbol{\Theta}\big\{(\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\Gamma}}_k)^\top(\mathbf{Y}_k - \mathbf{X}_k\hat{\boldsymbol{\Gamma}}_k)/n_k - \xi_k^\top\xi_k/n_k\big\}\boldsymbol{\Theta}||_\infty.$$

Adding and subtracting $\mathbf{X}_k\boldsymbol{\Gamma}$ to the first term in $\mathbf{R}^4_{k,\boldsymbol{\Theta}}$, we get that

$$||\mathbf{R}^4_{k,\boldsymbol{\Theta}}||_\infty \leq |||\boldsymbol{\Theta}|||^2_\infty\bigg\{2||(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})^\top\mathbf{X}_k^\top\xi_k||_\infty/n_k + ||\mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})||^2_F/n_k\bigg\}.$$

Under assumption (B1) and by conditioning on the event $\mathcal{F}_k(n_k, p, q)$, we have

$$||\mathbf{R}^4_{k,\boldsymbol{\Theta}}||_\infty \leq d_1\big\{\rho_k||\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}||_1 + ||\mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})||^2_F/n_k\big\}.$$

Recall that S_2 and $S_2(b)$ denote the support of the coefficient matrix $\boldsymbol{\Gamma}$ and its b-th column, $b = 1, \ldots, p$, with cardinalities s_2 and $s_2(b)$, respectively. Under condition (46), by a similar argument as in the proof of Lemma 3, it holds that for the fixed design matrix $\mathbf{X}_k$,

$$||\mathbf{X}_k(\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma})||^2_F/n_k \geq \mu^2_{\min}||\hat{\boldsymbol{\Gamma}}_k - \boldsymbol{\Gamma}||^2_F,$$

where $\mu_{\min} = \min_{1\leq b\leq p} \mu_b$. The reader can refer to Lemmas 3 and 4 and their preliminaries for more details. Combining this inequality with (55) and (56) from the proof of Lemma 4, we get

$$2\rho_k||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}||_1 + ||\mathbf{X}_k(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})||_F^2/n_k \leq 16\rho_k^2 s_2/\mu_{\min}^2.$$

Since $\mu_{\min} \in (0, \infty)$, there exists $L = O(1)$ such that $1/\mu_{\min}^2 \leq L$. Combining this with $\rho_k \asymp \sqrt{\log(pq)/n_k}$, and substituting in $\mathbf{R}_{k,\mathbf{\Theta}}^4$, we get

$$||\mathbf{R}_{k,\mathbf{\Theta}}^4||_\infty = O_p(d_1 s_2 \log(pq)/n_k).$$

Substituting the obtained bounds in (38) and considering $\kappa_{\mathbf{\Sigma}} = O(1)$, $\kappa_{\mathbf{H}} = O(1)$, $1/\alpha_1 = O(1)$, and $\gamma = O(1)$, we get

$$||\mathbf{R}_{k,\mathbf{\Theta}}||_\infty = O_p\bigg(\sqrt{n_k}\max\big\{d_1^{3/2}\log(p)/n_k, d_1^2(\log(p)/n_k)^{3/2}, d_1\log(p)/n_k, \\ d_1 s_2 \log(pq)/n_k\big\}\bigg),$$

and under the sparsity assumptions $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$ and $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$ the result $||\mathbf{R}_{k,\mathbf{\Theta}}||_\infty = o_p(1)$ follows. □

A.4 Proof of Theorem 3

The proof of this theorem is an extension of the proof of Theorem 1 from Jankova and van de Geer (2015) by considering a non-zero mean structure. Under the assumptions of Lemma 1 from the main text, it is shown that $||\mathbf{R}_{k,\mathbf{\Theta}}||_\infty = o_p(1)$. As such, using (10), we have

$$\begin{aligned}\sqrt{n_k}(\hat{\mathbf{\Theta}}_{ab,k}^d - \mathbf{\Theta}_{ab}) &= -\sqrt{n_k}\{\mathbf{\Theta}\mathbf{W}_k\mathbf{\Theta}\}_{ab} + o_p(1) \\ &= -\frac{1}{\sqrt{n_k}}\sum_{l=1}^{n_k}\big(\mathbf{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \mathbf{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \mathbf{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\mathbf{\Theta}_b \\ &\quad - \mathbf{\Theta}_{ab}\big) + o_p(1), \end{aligned} \tag{39}$$

where $\dot{\mathbf{Y}}_{l,k} \in \mathbb{R}^p$ and $\dot{\mathbf{X}}_{l,k} \in \mathbb{R}^q$ are the l-th row of the sub-matrices $\mathbf{Y}_k$ and $\mathbf{X}_k$, respectively, and $\mathbf{\Theta}_a$ and $\mathbf{\Theta}_b$ are p-dimensional vectors coming from the a-th row and the b-th row of $\mathbf{\Theta}$. The second equality in (39) follows directly since $\mathbf{\Theta}\mathbf{W}_k\mathbf{\Theta} = \mathbf{\Theta}\hat{\mathbf{\Sigma}}_{k,\mathbf{\Gamma}}\mathbf{\Theta} - \mathbf{\Theta}$. To show the normality of the term in (39),

define $\mathbf{Z}_{ab,l,k} := \boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}$, $l = 1, \ldots, n_k$. Then,

$$\begin{aligned}\mathbb{E}(\mathbf{Z}_{ab,l,k}) &= \boldsymbol{\Theta}_a^\top\mathbb{E}\big\{(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\big\}\boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab} \\ &= \mathbf{e}_a^\top\boldsymbol{\Theta}\mathbf{e}_b - \boldsymbol{\Theta}_{ab} = 0,\end{aligned}$$

where $\mathbf{e}_b$ is a p-dimensional unit column vector of zeros with one at position b and similarly for $\mathbf{e}_a$. On the other hand, since $\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k}$ is a Gaussian random vector, the variance

$$\sigma_{ab}^2 = \mathbb{V}\mathrm{ar}(\mathbf{Z}_{ab,l,k}) = \mathbb{V}\mathrm{ar}\big(\boldsymbol{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\boldsymbol{\Theta}_b\big)$$

is finite. Denote by $\mathcal{Z}_{n_k} = \sum_{l=1}^{n_k}\mathbf{Z}_{ab,l,k}$. Then, $z_{n_k}^2 := \mathbb{V}\mathrm{ar}(\mathcal{Z}_{n_k}) = n\sigma_{ab}^2$. Dividing (39) by $\sigma_{ab} > 0$, we get

$$\sqrt{n_k}(\hat{\boldsymbol{\Theta}}_{ab,k}^d - \boldsymbol{\Theta}_{ab})/\sigma_{ab} = \mathcal{Z}_{n_k}/z_{n_k} + o_p(1)/\sigma_{ab}.$$

It is enough to show that $\mathcal{Z}_{n_k}/z_{n_k} \xrightarrow{d} \mathcal{N}(0,1)$, where $\xrightarrow{d}$ denotes convergence in distribution. By substituting $\mathbf{Z}_{ab,l,k}$ in the proof of Theorem 1 from Jankova and van de Geer (2015), with the same argument, the normality of $\mathcal{Z}_{n_k}/z_{n_k}$ follows. The reader can refer to Jankova and van de Geer (2015) for more details.

To show $\sigma_{ab}^2 = \boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb} + \boldsymbol{\Theta}_{ab}^2$, under the multivariate Gaussian distribution of $\dot{\varepsilon}_{l,k}$, which is the l-th row of ξ_k, we have $\dot{\varepsilon}_{l,k} = \dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, and as such $\boldsymbol{\Theta}(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top\dot{\mathbf{X}}_{l,k}) \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Theta})$. Using a similar argument as in the proof of Lemma 2 from Jankova and van de Geer (2015), it holds that $\sigma_{ab}^2 = \boldsymbol{\Theta}_{aa}\boldsymbol{\Theta}_{bb} + \boldsymbol{\Theta}_{ab}^2$ and $1/\sigma_{ab} = O(1)$. □

A.5 Proof of Theorem 4

a) The proof of this theorem relies on Lemma 6, which shows the negligibility of the remainder term $\mathbf{R}_{a,\boldsymbol{\Gamma}}$ as $n_\dagger$ and K grow. To show the asymptotic normality of $\mathbf{W}_{a,\boldsymbol{\Gamma}}$, define

$$\zeta_a = \sum_{k=1}^K \sqrt{\frac{n_k[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{n[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}},$$

where $\mathbf{T}_{a,k}$ is the a-th element of $\mathbf{T}_k$ from (5). As $\sqrt{\frac{\sum_{k=1}^K n_k/[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^K n_k/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}} \xrightarrow{p} 1$, see (66), using Slutsky's theorem, it is enough to show that ζ_a converges

in distribution to $\mathcal{N}(0,1)$. Defining $\zeta'_a = \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}}$, we show in Lemma 7 that $|\zeta_a - \zeta'_a| \xrightarrow{p} 0$ as $K \to \infty$ and $n_k \to \infty$, $k = 1, \ldots, K$, and then convergence in distribution of ζ_a follows by convergence in distribution of ζ'_a.

Now to show the asymptotic normal distribution of ζ'_a, denoting by $\mathbf{Z}_{a,k} := \sqrt{\frac{n_k}{n}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}}$, we use the Lindeberg theorem (see for instance, Theorem 2.1 of Gut, 2005). We have that

$$\begin{aligned}
\mathbb{E}(\mathbf{Z}_{a,k}) &= \mathbb{E}\bigg\{\mathbb{E}\bigg(\sqrt{\frac{n_k}{n}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}} \mid \mathbf{X}_k\bigg)\bigg\} \\
&= \mathbb{E}\bigg\{\sqrt{\frac{n_k}{n}} \times \frac{1}{\sqrt{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}} \mathbb{E}\big(\mathbf{T}_{a,k} \mid \mathbf{X}_k\big)\bigg\} = 0,
\end{aligned}$$

where the last equality is deduced from the conditional normal distribution of $\mathbf{T}_{a,k}$ shown in Theorem 1. Moreover,

$$\sum_{k=1}^{K} \mathbb{E}(\mathbf{Z}_{a,k}^2) = \sum_{k=1}^{K} \mathbb{E}\bigg\{\frac{n_k}{n} \times \frac{1}{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} \mathbb{E}\big(\mathbf{T}_{a,k}^2 \mid \mathbf{X}_k\big)\bigg\} = \sum_{k=1}^{K} \frac{n_k}{n} = 1,$$

where the second equality is deduced using the conditional variance of $\mathbf{T}_{a,k}$, which is equal to $[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}$. Now to check the Lindeberg condition, for every $\epsilon > 0$, we can write

$$\begin{aligned}
&\mathbb{E}\big(\mathbf{Z}_{a,k}^2 \mathbb{I}(|\mathbf{Z}_{a,k}| > \epsilon) \mid \mathbf{X}_k\big) \\
&\quad = \frac{n_k}{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} \mathbb{E}\bigg\{\mathbf{T}_{a,k}^2 \mathbb{I}\big(|\mathbf{T}_{a,k}| > \epsilon \sqrt{\frac{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}{n_k}}\big) \mid \mathbf{X}_k\bigg\},
\end{aligned} \tag{40}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Following Jankova and van de Geer (2015), pp. 1223, for a random variable X and a positive constant a, one can write

$$\mathbb{E}\big(X^2 \mathbb{I}(|X| > a)\big) = a^2 \mathbb{P}(|X| > a) + 2 \int_a^\infty u \mathbb{P}(|X| > u)\, du.$$

Applying this equality to (40), we get

$$\begin{aligned}
&\mathbb{E}\big(\mathbf{Z}_{a,k}^2 \mathbb{I}(|\mathbf{Z}_{a,k}| > \epsilon) \mid \mathbf{X}_k\big) \\
&\quad = \epsilon^2 \mathbb{P}\Big(|\mathbf{T}_{a,k}| > \epsilon \sqrt{\frac{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}{n_k}} \mid \mathbf{X}_k\Big) \\
&\quad + \frac{2n_k}{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} \int_{\epsilon \sqrt{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}/n_k}}^{\infty} u \mathbb{P}\big(|\mathbf{T}_{a,k}| > u \mid \mathbf{X}_k\big)\, du.
\end{aligned} \tag{41}$$

Applying the concentration inequality $\mathbb{P}(|X - \mu| > t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$, $t \in \mathbb{R}$, for a normal random variable X with mean μ and variance σ^2, we get

$$\mathbb{P}\left(|\mathbf{T}_{a,k}| > \epsilon \sqrt{\frac{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}{n_k}} \mid \mathbf{X}_k\right) \leq 2e^{-\frac{n\epsilon^2}{2n_k}}.$$

In the integral, by substituting $t := \frac{\sqrt{n_k} u}{\epsilon \sqrt{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}}$, we have

$$\begin{aligned}
&\int_{\epsilon \sqrt{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}/n_k}}^{\infty} u \mathbb{P}\big(|\mathbf{T}_{a,k}| > u \mid \mathbf{X}_k\big)\, du \\
&= \frac{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa} \epsilon^2}{n_k} \int_1^{\infty} t \mathbb{P}\left(|\mathbf{T}_{a,k}| > \epsilon t \sqrt{\frac{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}{n_k}} \mid \mathbf{X}_k\right) dt \\
&\leq \frac{n[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa} \epsilon^2}{n_k} \int_1^{\infty} 2te^{-\frac{n\epsilon^2 t^2}{2n_k}}\, dt = 2[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa} e^{-\frac{n\epsilon^2}{2n_k}}.
\end{aligned}$$

As such, in (41),

$$\mathbb{E}\big(\mathbf{Z}_{a,k}^2 \mathbb{I}(|\mathbf{Z}_{a,k}| > \epsilon) \mid \mathbf{X}_k\big) \leq 2e^{-\frac{n\epsilon^2}{2n_k}} \{\epsilon^2 + 2n_k/n\}.$$

Thus, in the Lindeberg condition, we get

$$\begin{aligned}
&\lim_{K\to\infty}\lim_{\substack{n_k\to\infty\\ k=1,\ldots,K}}\sum_{k=1}^{K}\mathbb{E}\Big\{\mathbf{Z}_{a,k}^2\mathbb{I}\big(|\mathbf{Z}_{a,k}|>\epsilon\big)\Big\}\\
&\quad=\lim_{K\to\infty}\lim_{\substack{n_k\to\infty\\ k=1,\ldots,K}}\sum_{k=1}^{K}\mathbb{E}\Big\{\mathbb{E}\big(\mathbf{Z}_{a,k}^2\mathbb{I}(|\mathbf{Z}_{a,k}|>\epsilon)\mid\mathbf{X}_k\big)\Big\}\\
&\quad\le\lim_{K\to\infty}\lim_{\substack{n_k\to\infty\\ k=1,\ldots,K}}\sum_{k=1}^{K}2e^{-\frac{n\epsilon^2}{2n_k}}\{\epsilon^2+2n_k/n\}\\
&\quad\le\lim_{K\to\infty}\lim_{\substack{n_k\to\infty\\ k=1,\ldots,K}}\sum_{k=1}^{K}2e^{-\frac{Kn_\dagger\epsilon^2}{2n}}\{\epsilon^2+2n_k/n\}=0,
\end{aligned}$$

where the last equality follows by the dominated convergence theorem.

b) By basic algebra it can be shown that

$$\frac{\sqrt{nK}}{\sqrt{\sum_{k=1}^{K}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\big(\mathrm{vec}(\tilde{\boldsymbol{\Gamma}}_{\mathrm{wAvg}})_a-\mathrm{vec}(\boldsymbol{\Gamma})_a\big)=\mathbf{W}_{a,\boldsymbol{\Gamma}}^{'}+\mathbf{R}_{a,\boldsymbol{\Gamma}}^{'},$$

where

$$\begin{aligned}
\mathbf{W}_{a,\boldsymbol{\Gamma}}^{'}&=\sqrt{\frac{K[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\sum_{k=1}^{K}\sqrt{\frac{n_k}{n}}\frac{1}{\sqrt{[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}\mathbf{T}_{a,k},\\
\mathbf{R}_{a,\boldsymbol{\Gamma}}^{'}&=\sqrt{\frac{K[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\sum_{k=1}^{K}\sqrt{\frac{n_k}{n}}\frac{1}{\sqrt{[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}\mathbf{R}_{a,k,\boldsymbol{\Gamma}},
\end{aligned}$$

where $\mathbf{T}_{a,k}$ and $\mathbf{R}_{a,k,\boldsymbol{\Gamma}}$ are respectively the a-th element of $\mathbf{T}_k$ and $\mathbf{R}_{k,\boldsymbol{\Gamma}}$ defined in (5). Consider the random variable $\zeta_a=\frac{1}{\sqrt{[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}\sum_{k=1}^{K}\sqrt{\frac{n_k}{n}}\mathbf{T}_{a,k}$. In part a), it is shown that $\zeta_a'=\sum_{k=1}^{K}\sqrt{\frac{n_k}{n}}\times\frac{1}{\sqrt{[\boldsymbol{\Sigma}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\mathbf{T}_{a,k}$ converges in distribution to $\mathcal{N}(0,1)$. On the other hand, one can easily show that $|\zeta_a-\zeta_a'|=O_p\big(Kd_2\sqrt{\log(pq)\log(q)/n}\big)$, and then the $o_p(1)$ result follows by the mentioned sparsity condition on d_2 and with $K=O\big(n^{\pi_5}/(\sqrt{\log(pq)\log(q)}\ \max\{s_2,d_2\})\big)$, $0<\pi_5<1/2$. As such, the asymptotic normality of ζ_a follows directly. Moreover, the $o_p(1)$ result of the remainder term $\mathbf{R}_{a,\boldsymbol{\Gamma}}^{'}$ is reached by the same technique as in part a).

c) By basic algebra it can be shown that

$$\frac{K^{3/2}}{\sqrt{(\sum_{k=1}^{K}[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa})(\sum_{k=1}^{K}1/n_k)}}\big(\mathrm{vec}(\tilde{\mathbf{\Gamma}}_{\mathrm{sAvg}})_a-\mathrm{vec}(\mathbf{\Gamma})_a\big)$$
$$=\mathbf{W}''_{a,\mathbf{\Gamma}}+\mathbf{R}''_{a,\mathbf{\Gamma}},$$

where

$$\mathbf{W}''_{a,\mathbf{\Gamma}}=\frac{\sqrt{K[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}{\sqrt{\sum_{k=1}^{K}[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\times\frac{1}{\sqrt{([\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa})(\sum_{k=1}^{K}1/n_k)}}$$
$$\times\sum_{k=1}^{K}\frac{1}{\sqrt{n_k}}\mathbf{T}_{a,k},$$
$$\mathbf{R}''_{a,\mathbf{\Gamma}}=\frac{\sqrt{K[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}{\sqrt{\sum_{k=1}^{K}[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}\times\frac{1}{\sqrt{([\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa})(\sum_{k=1}^{K}1/n_k)}}.$$
$$\times\sum_{k=1}^{K}\frac{1}{\sqrt{n_k}}\mathbf{R}_{a,k,\mathbf{\Gamma}}.$$

Considering $\zeta_a=\frac{1}{\sqrt{([\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa})(\sum_{k=1}^{K}1/n_k)}}\sum_{k=1}^{K}\frac{1}{\sqrt{n_k}}\mathbf{T}_{a,k}$ and $\zeta'_a=\sum_{k=1}^{K}\frac{1}{\sqrt{n_k[\mathbf{\Sigma}\otimes(\mathbf{M_k}\mathbf{C_k}\mathbf{M_k}^\top)]_{aa}(\sum_{k=1}^{K}1/n_k)}}\mathbf{T}_{a,k}$, by similar techniques as in part a), using the Lindeberg theorem, one can show that ζ'_a is asymptotically normal with mean zero and variance 1. Moreover, $|\zeta_a-\zeta'_a|=O_p\big(d_2\sqrt{nK\log(pq)\log(q)}/n_\dagger\big)$ and $|\mathbf{R}''_{a,\mathbf{\Gamma}}|=O_p\big(s_2\sqrt{nK\log(pq)\log(q)}/n_\dagger\big)$, which are $o_p(1)$ under the mentioned sparsity conditions and with $\sqrt{K}=O\big(n_\dagger^{\pi_6}/(\sqrt{\log(pq)\log(q)}\ \max\{s_2,d_2\})\big)$, $0<\pi_6<1/2$. □

A.6 Proof of Theorem 5

a) The proof of this theorem relies on Lemma 8, which shows the negligibility of the remainder term $\mathbf{R}_{ab,\mathbf{\Theta}}$ as $n_\dagger$ and K both grow. To show the asymptotic normality, define

$$\zeta_{ab}=\sum_{k=1}^{K}\frac{1}{\sqrt{n}\sigma_{ab}}\times\frac{\sigma_{ab}^2}{\hat{\sigma}_{ab,k}^2}\sum_{l=1}^{n_k}\big(\mathbf{\Theta}_a^\top(\dot{\mathbf{Y}}_{l,k}-\mathbf{\Gamma}^\top\dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k}-\mathbf{\Gamma}^\top\dot{\mathbf{X}}_{l,k})^\top\mathbf{\Theta}_b-\mathbf{\Theta}_{ab}\big),$$

where $\dot{\mathbf{Y}}_{l,k}$ and $\dot{\mathbf{X}}_{l,k}$ are the l-th row, $l = 1, \ldots, n_k$, of $\mathbf{Y}_k$ and $\mathbf{X}_k$, respectively. Similarly to the proof of Theorem 4, by defining the sequence

$$\zeta'_{ab} = \sum_{k=1}^{K} \frac{1}{\sqrt{n}\sigma_{ab}} \sum_{l=1}^{n_k} \big(\boldsymbol{\Theta}_a^\top (\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k})^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}\big),$$

we first show that $|\zeta_{ab} - \zeta'_{ab}| \xrightarrow{p} 0$ as $K \to \infty$ and $n_k \to \infty$, $k = 1, \ldots, K$, and then convergence in distribution of ζ'_{ab} yields convergence in distribution of ζ_{ab}. As it is shown in the proof of Lemma 8, the term $(1/\hat{\sigma}^2_{ab,k}) = O_p(1)$. Using Remark 4, we have

$$\begin{aligned} \big|\zeta_{ab} - \zeta'_{ab}\big| \leq \sum_{k=1}^{K} \frac{1}{\sqrt{n}\sigma_{ab}} \times O_p\big(\max\{\log(p)/n_k, \sqrt{d_1 \log(p)/n_k}\}\big) \\ \times \left| \sum_{l=1}^{n_k} \big(\boldsymbol{\Theta}_a^\top (\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k})^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}\big) \right|. \quad (42) \end{aligned}$$

Due to the definition of $\mathbf{W}_k$ in Section 4, we have that $\big|\sum_{l=1}^{n_k} \big(\boldsymbol{\Theta}_a^\top (\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k})^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}\big)\big| = n_k \big|\{\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}\}_{ab}\big|$. Under assumption (B2) and using Lemma 2 of Lam and Fan (2009), we have

$$n_k \big|\{\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}\}_{ab}\big| \leq n_k ||\boldsymbol{\Theta}\mathbf{W}_k\boldsymbol{\Theta}||_\infty \leq n_k |||\boldsymbol{\Theta}|||_\infty^2 ||\mathbf{W}_k||_\infty \leq d_1 \sqrt{n_k \log(p)}.$$

Substituting this bound in (42), we get

$$|\zeta_{ab} - \zeta'_{ab}| \leq O_p\big(\frac{K}{\sqrt{n}} \max\big\{d_1 (\log(p))^{3/2}/\sqrt{n_\dagger}, d_1^{3/2} \log(p)\big\}\big),$$

and under the conditions $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$ and $K = O\big(n^{\pi_7}/(d_1 \log(p))\big)$, $0 < \pi_7 \leq 1/3$, the result $|\zeta_{ab} - \zeta'_{ab}| = o_p(1)$ follows.

Now to show the asymptotic normality of ζ'_{ab} we use the Lindeberg-Feller theorem (see for instance, Theorem 4.12 of Kallenberg, 1997). By defining $\mathbf{Z}_{ab,l,k} := \boldsymbol{\Theta}_a^\top (\dot{\mathbf{Y}}_{l,k}$
$- \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k})(\dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k})^\top \boldsymbol{\Theta}_b - \boldsymbol{\Theta}_{ab}$, we have $\mathbb{E}(\mathbf{Z}_{ab,l,k}/(\sqrt{n}\sigma_{ab})) = 0$. Moreover,

$$\lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \sum_{l=1}^{n_k} \mathbb{E}(\mathbf{Z}^2_{ab,l,k}/(n\sigma^2_{ab})) = \lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \frac{n_k}{n} = \lim_{K\to\infty} \sum_{k=1}^{K} c_k = 1,$$

where the first equality is deduced based on the definition of $\mathbb{V}\text{ar}(\mathbf{Z}_{ab,l,k})$ which is equal to σ^2_{ab} under the p-dimensional Gaussian distribution of $\dot{\varepsilon}_{l,k} = \dot{\mathbf{Y}}_{l,k} - \boldsymbol{\Gamma}^\top \dot{\mathbf{X}}_{l,k}$. Since the term n_k/n is bounded and $\lim_{n_k\to\infty} n_k/n = c_k$, using the dominated convergence theorem, the second equality also follows. Since $\mathbf{Z}_{ab,l,k}$ are identically distributed for any $l = 1, \ldots, n_k$ and $k = 1, \ldots, K$, we can write

$$\begin{aligned}
&\lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \sum_{l=1}^{n_k} \frac{1}{n\sigma^2_{ab}} \mathbb{E}\big(\mathbf{Z}^2_{ab,l,k} \mathbb{I}(|\mathbf{Z}_{ab,l,k}| > \sqrt{n}\sigma_{ab}\epsilon)\big) \\
&= \lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \frac{n_k}{n\sigma^2_{ab}} \mathbb{E}\big(\mathbf{Z}^2_{ab,1,1} \mathbb{I}(|\mathbf{Z}_{ab,1,1}| > \sqrt{n}\sigma_{ab}\epsilon)\big) \\
&= \left(\lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \frac{1}{\sigma^2_{ab}} \mathbb{E}\big(\mathbf{Z}^2_{ab,1,1} \mathbb{I}(|\mathbf{Z}_{ab,1,1}| > \sqrt{n}\sigma_{ab}\epsilon)\big) \right) \\
&\quad \times \left(\lim_{K\to\infty} \lim_{\substack{n_k\to\infty \\ k=1,\ldots,K}} \sum_{k=1}^{K} \frac{n_k}{n} \right). \qquad (43)
\end{aligned}$$

Under the sparsity condition $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$, Jankova and van de Geer (2015) showed that $\lim_{n\to\infty} \frac{1}{\sigma^2_{ab}} \times \mathbb{E}\big(\mathbf{Z}^2_{ab,1,1} \mathbb{I}(|\mathbf{Z}_{ab,1,1}| > \sqrt{n}\sigma_{ab}\epsilon)\big) = 0$. Due to the fact that $n = \sum_{k=1}^{K} n_k$, the first limit in (43) is zero and using the dominated convergence theorem, the second limit is equal to 1. As such, the requirements of the Lindeberg-Feller condition hold and ζ'_{ab} converges in distribution to $\mathcal{N}(0, 1)$.

To show the results of parts b) and c), we refer the reader to part b) and c) of Appendix A.5 due to the similar technique which can be used here as well. □

B Some technical lemmas and their proofs

Lemma 2. Consider the regression model (2) with random noise matrix ξ_k and random Gaussian design matrix $\mathbf{X}_k$. Supposing $[\mathbf{C}_k]_{aa} = 1$, $a = 1, \ldots, q$, where $[\mathbf{C}_k]_{aa}$ is the a-th diagonal element of $\mathbf{C}_k = \mathbf{X}_k^\top \mathbf{X}_k/n_k$, and choosing $\rho_0 = (\max_{1\le b\le p} \sqrt{\boldsymbol{\Sigma}_{bb}}) \sqrt{A \log(pq)/n_k}$, where $A > 2$ is a constant, we have

$$\mathbb{P}\big(\mathcal{F}_k(n_k, p, q)\big) \ge 1 - 2/(pq)^{A/2-1}.$$

Proof. According to the definition of the elementwise ℓ_∞ norm, the event $\mathcal{F}_k$ is equivalent to

$$\mathcal{F}_k(n_k, p, q) = \Big\{ \max_{1\le a\le q} \max_{1\le b\le p} \big| \sum_{l=1}^{n_k} \varepsilon_{l,k}^b X_{l,k}^a \big| / n_k \le \rho_0 \Big\},$$

where $X_{l,k}^a$ and $\varepsilon_{l,k}^b$ are the a-th and the b-th components of the vectors $\dot{\mathbf{X}}_{l,k}$ and $\dot{\varepsilon}_{l,k}$ which are the l-th row of $\mathbf{X}_k$ and ξ_k, respectively. Conditioning on $X_{1,k}^a, \ldots, X_{n_k,k}^a$, the random variable $\sum_{l=1}^{n_k} \varepsilon_{l,k}^b X_{l,k}^a$ is a linear transformation of $\varepsilon_{1,k}^b, \ldots, \varepsilon_{n_k,k}^b$ and it is normally distributed with mean zero and variance $n_k \boldsymbol{\Sigma}_{bb} [\mathbf{C}_k]_{aa}$. Supposing $[\mathbf{C}_k]_{aa} = 1$, we get $V_{ab} \mid (X_{1,k}^a, \ldots, X_{n_k,k}^a)^\top \sim \mathcal{N}(0,1)$, where $V_{ab} := \sum_{l=1}^{n_k} \varepsilon_{l,k}^b X_{l,k}^a / \sqrt{n_k \boldsymbol{\Sigma}_{bb}}$. With our choice of ρ_0, we have that,

$$\begin{aligned}
\mathbb{P}\Big(\max_{\substack{1\le a\le q \\ 1\le b\le p}} \big| \sum_{l=1}^{n_k} \varepsilon_{l,k}^b X_{l,k}^a \big| / n_k \ge \rho_0 \Big) &= \mathbb{P}\Big(\max_{1\le a\le q} \frac{\max_{1\le b\le p} \big| \sum_{l=1}^{n_k} \varepsilon_{l,k}^b X_{l,k}^a \big|}{n_k \max_{1\le b\le p} \sqrt{\boldsymbol{\Sigma}_{bb}}} \\
&\qquad > \sqrt{A \log(pq)/n_k} \Big) \\
&\le \mathbb{P}\Big(\max_{\substack{1\le a\le q \\ 1\le b\le p}} \Big| \frac{\sum_{l=1}^{n_k} \varepsilon_{l,k}^b X_{l,k}^a}{\sqrt{n_k \boldsymbol{\Sigma}_{bb}}} \Big| > \sqrt{A \log(pq)} \Big) \\
&\le \sum_{a=1}^{q} \sum_{b=1}^{p} \mathbb{P}\big(|V_{ab}| \ge \sqrt{A \log(pq)} \big). \qquad (44)
\end{aligned}$$

To obtain a bound on $\mathbb{P}\big(|V_{ab}| \ge \sqrt{A \log(pq)}\big)$, by conditioning on $\mathbf{X}^a := (X_{1,k}^a, \ldots, X_{n_k,k}^a)^\top$, $a = 1, \ldots, q$, and the Gaussian concentration inequality $\mathbb{P}(|X - \mu| > \sigma t) \le 2\exp(-t^2/2)$, $t \in \mathbb{R}$, for a normal random variable X with mean μ and variance σ^2, we get

$$\begin{aligned}
\mathbb{P}\big(|V_{ab}| \ge \sqrt{A \log(pq)} \big) &= \int \mathbb{P}\big(|V_{ab}| \ge \sqrt{A \log(pq)} \mid \mathbf{X}^a = \mathbf{x}^a \big) f_{\mathbf{X}^a}(\mathbf{x}^a)\, d\mathbf{x}^a \\
&\le 2\exp\{-A \log(pq)/2\} \int f_{\mathbf{X}^a}(\mathbf{x}^a)\, d\mathbf{x}^a = \frac{2}{(pq)^{A/2}}. \qquad (45)
\end{aligned}$$

Substituting (45) in (44), the result of lemma follows directly. □

Before providing Lemmas 3 and 4, we need some preliminary notation. The results are provided under the Restricted Eigenvalue (RE) condition. The following definition from Raskutti et al. (2010) defines the RE condition for the population covariance matrix and is essentially equivalent to the RE condition of Bickel et al. (2009). For a given subset $G \subset \{1, \ldots, q\}$ with cardinality $g = \#G$ and a constant $\delta \geq 1$, define the set

$$C(G; \delta) := \{\nu \in \mathbb{R}^q \mid ||\nu_{G^c}||_1 \leq \delta ||\nu_G||_1\},$$

where ν_G is the restriction of ν to G and has zeros outside the set G. The symbol G^c denotes the complement set of G.

Definition 1. (Raskutti et al., 2010) A deterministic covariance matrix $\mathbf{Q}$ satisfies the RE condition over G of order g, if there exist constants $(\delta, \mu) \in [1, \infty) \times (0, \infty)$ such that

$$||\mathbf{Q}^{1/2}\nu||_2 \geq \mu ||\nu||_2, \quad \text{for all } \nu \in C(G; \delta),$$

where $\mathbf{Q}^{1/2}$ is the square root matrix of $\mathbf{Q}$.

Note that Definition 1 provides a lower bound on the ℓ_2 norm of the product between $\mathbf{Q}^{1/2}$ and the vector ν. As in the multivariate regression we have a matrix of coefficients not a vector, we need to provide a lower bound on the product of $\mathbf{Q}^{1/2}$ and each column of the coefficient matrix. To this end, consider the matrix $\mathbf{V} \in \mathbb{R}^{q \times p}$ in general, and denote its vectorized form with $\nu := \text{vec}(\mathbf{V})$ such that $\nu = (\nu_{(1)}^\top, \ldots, \nu_{(p)}^\top)^\top \in \mathbb{R}^{qp}$, where $\nu_{(b)} \in \mathbb{R}^q$ is the b-th column of $\mathbf{V}$, $b = 1, \ldots, p$. To simplify the notation and having one index for the elements of ν, denote the index set of $\nu_{(b)}$ as $\{q(b-1)+1, \ldots, qb\}$, $b = 1, \ldots, p$. Consider $G_b \subset \{q(b-1)+1, \ldots, qb\}$ with cardinality $g_b = \#G_b$ and complement set G_b^c. For a constant $\delta_b \geq 1$, define the set

$$C(G_b; \delta_b) := \left\{\nu_{(b)} \in \mathbb{R}^q \mid ||[\nu_{(b)}]_{G_b^c}||_1 \leq \delta_b ||[\nu_{(b)}]_{G_b}||_1\right\},$$

where $[\nu_{(b)}]_{G_b}$ is the restriction of $\nu_{(b)}$ to G_b which has zeros outside the subset G_b.

Definition 2. A deterministic covariance matrix $\mathbf{Q}$ satisfies the multivariate restricted eigenvalue (MRE) condition over $G = \{G_1, \ldots, G_p\}$ of order $g = \sum_{b=1}^p g_b$, if there exist constants $(\delta_b, \mu_b) \in [1, \infty) \times (0, \infty)$, such that for every $b = 1, \ldots, p$,

$$||\mathbf{Q}^{1/2}\nu_{(b)}||_2 \geq \mu_b ||\nu_{(b)}||_2, \quad \text{for all} \quad \nu_{(b)} \in C(G_b; \delta_b).$$

Definition 3. The sample covariance matrix $\mathbf{C}_k = \mathbf{X}_k^\top \mathbf{X}_k / n_k$ of the design $\mathbf{X}_k$ satisfies the MRE condition over $G = \{G_1, \ldots, G_p\}$ of order $g = \sum_{b=1}^p g_b$, if there exist constants $(\delta_b, \mu_b) \in [1, \infty) \times (0, \infty)$, such that for every $b = 1, \ldots, p$,

$$\nu_{(b)}^\top \mathbf{C}_k \nu_{(b)} \equiv ||\mathbf{X}_k \nu_{(b)}||_2^2 / n_k \geq \mu_b^2 ||\nu_{(b)}||_2^2, \quad \text{for all} \quad \nu_{(b)} \in C(G_b; \delta_b). \tag{46}$$

To introduce Lemma 3, let $S_2(b)$ be the support of the b-th column of $\mathbf{\Gamma}$. By vectorizing $\mathbf{\Gamma}$ and rewriting its columns as explained for the general matrix $\mathbf{V}$, we have $S_2(b) \subset \{q(b-1)+1, \ldots, qb\}$, with cardinality $s_2(b) = \# S_2(b)$ and complement set $S_2^c(b)$. For a constant $\delta_b \geq 1$, denote the set

$$C(S_2(b); \delta_b) := \big\{ \nu_{(b)} \in \mathbb{R}^q \mid ||[\nu_{(b)}]_{S_2^c(b)}||_1 \leq \delta_b ||[\nu_{(b)}]_{S_2(b)}||_1 \big\},$$

where $[\nu_{(b)}]_{S_2(b)}$ is the restriction of $\nu_{(b)}$ to $S_2(b)$ which has zeros outside the subset $S_2(b)$.

Lemma 3. Consider the regression model (2) for the k-th sub-sample with random Gaussian design $\mathbf{X}_k$ which has independent and identical $\mathcal{N}_q(\mathbf{0}, \mathbf{Q})$ rows, and denote the maximum diagonal element of $\mathbf{Q}$ by $\mathbf{Q}_{\max}$. Suppose that $\mathbf{Q}$ satisfies the MRE condition over S_2 of order s_2 for all vectors in $C(S_2(b); \delta_b)$ with parameters (δ_b, μ_b), $b = 1, \ldots, p$. On the event $\mathcal{F}_k(n_k, p, q)$, for some positive constants c, c' and c'', if the sub-sample size n_k, $k = 1, \ldots, K$, satisfies

$$n_k \geq c'' \frac{\mathbf{Q}_{\max}^2 (1 + \delta_{\max})^2}{\mu_{\min}^2} s_2 \log(q), \tag{47}$$

where $\mu_{\min} = \min_{1 \leq b \leq p} \mu_b$ and $\delta_{\max} = \max_{1 \leq b \leq p} \delta_b$, then we have

$$\frac{||\mathbf{X}_k (\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})||_F^2}{n_k} \geq (\mu_{\min}/8)^2 ||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}||_F^2, \tag{48}$$

with probability at least $1 - c' \exp(-c n_k)$.

Proof. In general, consider the matrix $\mathbf{V} \in \mathbb{R}^{q \times p}$ with the b-th column $\nu_{(b)}$, $b = 1, \ldots, p$, and its vectorized form as $\nu \in \mathbb{R}^{qp}$. We have

$$\frac{||\mathbf{X}_k \mathbf{V}||_F}{\sqrt{n_k}} = \frac{||\mathrm{vec}(\mathbf{X}_k \mathbf{V})||_2}{\sqrt{n_k}} = \frac{||(\mathbf{I}_p \otimes \mathbf{X}_k)\nu||_2}{\sqrt{n_k}} = \sqrt{\frac{\sum_{b=1}^p ||\mathbf{X}_k \nu_{(b)}||_2^2}{n_k}}. \tag{49}$$

Using Theorem 1 of Raskutti et al. (2010), under the Gaussianity of the design matrix $\mathbf{X}_k$, we have

$$\frac{||\mathbf{X}_k\nu_{(b)}||_2}{\sqrt{n_k}} \geq \frac{1}{4}||\mathbf{Q}^{1/2}\nu_{(b)}||_2 - 9\mathbf{Q}_{\max}\sqrt{\frac{\log(q)}{n_k}}||\nu_{(b)}||_1, \quad \text{for all} \quad \nu_{(b)} \in \mathbb{R}^q,$$

with probability at least $1 - c'\exp(-cn_k)$. Under the MRE condition of $\mathbf{Q}$ for all vectors $\nu_{(b)} \in C(S_2(b); \delta_b)$ with parameters (μ_b, δ_b), $b = 1, \ldots, p$, and the relation between ℓ_1 and ℓ_2 norms, we get

$$\frac{||\mathbf{X}_k\nu_{(b)}||_2^2}{n_k} \geq \left\{\frac{1}{4}\mu_{\min} - 9\mathbf{Q}_{\max}(1+\delta_{\max})\sqrt{\frac{s_2\log(q)}{n_k}}\right\}^2 ||\nu_{(b)}||_2^2. \tag{50}$$

Substituting (50) in (49),

$$\frac{||\mathbf{X}_k\mathbf{V}||_F}{\sqrt{n_k}} \geq \left\{\frac{1}{4}\mu_{\min} - 9\mathbf{Q}_{\max}(1+\delta_{\max})\sqrt{\frac{s_2\log(q)}{n_k}}\right\}\sqrt{\sum_{b=1}^{p}||\nu_{(b)}||_2^2}. \tag{51}$$

Applying the lower bound (47) in (51) for some constant c'', we get

$$||\mathbf{X}_k\mathbf{V}||_F/\sqrt{n_k} \geq (\mu_{\min}/8)||\mathbf{V}||_F. \tag{52}$$

Note that using (55) in the proof of Lemma 4, we have

$$||\hat{\mathbf{\Gamma}}_{S_2^c,k} - \mathbf{\Gamma}_{S_2^c}||_1 \leq 4||\hat{\mathbf{\Gamma}}_{S_2,k} - \mathbf{\Gamma}_{S_2}||_1, \tag{53}$$

where $\mathbf{\Gamma}_{S_2}$ and $\hat{\mathbf{\Gamma}}_{S_2,k}$ are the matrices $\mathbf{\Gamma}$ and $\hat{\mathbf{\Gamma}}_k$ which have zeros outside the set S_2, respectively. By the same argument as in the proof of Lemma 4 for univariate linear regression (see for example, Bickel et al., 2009), a similar inequality to (53) can be written for the b-th column of $\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}$ and it can be deduced that the b-th column of $\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}$ is in $C(S_2(b); \delta_b)$. As such, by replacing $\mathbf{V}$ with $\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}$ in (52), the result in (48) follows directly. □

The following Lemma provides a convergence rate on the coefficient matrix estimator $\hat{\mathbf{\Gamma}}_k$, by a similar approach to that of Cai et al. (2013) for the Dantzig selector.

Lemma 4. Under the conditions of Lemma 3 and the lower bound (47) on the sub-sample size n_k, with probability at least $1 - c'\exp(-cn_k)$, we have

$$||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}||_1 = O_p\big(s_2\sqrt{\log(pq)/n_k}\big). \tag{54}$$

Proof. Consider the regression model (2) and recall the Lasso solution (3). Due to the fact that $\hat{\mathbf{\Gamma}}_k$ minimizes the loss in (3),

$$||\mathbf{X}_k(\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}_k)||_F^2/(2n_k) + \rho_k||\hat{\mathbf{\Gamma}}_k||_1 \leq \text{trace}\big((\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})^\top \mathbf{X}_k^\top \xi_k\big)/n_k + \rho_k||\mathbf{\Gamma}||_1,$$

where

$$\begin{aligned}
\text{trace}\big((\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})^\top \mathbf{X}_k^\top \xi_k\big) &\leq \Big|\text{trace}\big((\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})^\top \mathbf{X}_k^\top \xi_k\big)\Big| \\
&= \Big|\big(\text{vec}(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})\big)^\top \text{vec}(\mathbf{X}_k^\top \xi_k)\Big| \\
&\leq ||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}||_1 ||\text{vec}(\mathbf{X}_k^\top \xi_k)||_\infty,
\end{aligned}$$

where the last inequality holds due to Hölder's inequality. Now, on the event $\mathcal{F}_k(n_k, p, q)$, by a similar argument as in the univariate regression (see for example, chapter 6 of Bühlmann and Van De Geer, 2011), we get

$$\frac{||\mathbf{X}_k(\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}_k)||_F^2}{2n_k} + \frac{\rho_k||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}||_1}{2} \leq 2\rho_k||\hat{\mathbf{\Gamma}}_{S_2,k} - \mathbf{\Gamma}_{S_2}||_1, \tag{55}$$

where $\mathbf{\Gamma}_{S_2}$ and $\hat{\mathbf{\Gamma}}_{S_2,k}$ are the matrices $\mathbf{\Gamma}$ and $\hat{\mathbf{\Gamma}}_k$ which have zeros outside the set S_2, respectively. Due to the relation between the elementwise ℓ_1 and Frobenius norms of a matrix, we have

$$||\hat{\mathbf{\Gamma}}_{S_2,k} - \mathbf{\Gamma}_{S_2}||_1 \leq \sqrt{s_2}||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}||_F. \tag{56}$$

Substituting (56) in (55) and then applying Lemma 3 under the lower bound (47) on the sub-sample size n_k, we get

$$\frac{||\mathbf{X}_k(\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}_k)||_F^2}{2n_k} + \frac{\rho_k||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}||_1}{2} \leq 2\rho_k\sqrt{s_2}(\frac{8}{\mu_{\min}})\frac{||\mathbf{X}_k(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})||_F}{\sqrt{n_k}},$$

with probability at least $1 - c'\exp(-cn_k)$, and

$$||\mathbf{X}_k(\mathbf{\Gamma} - \hat{\mathbf{\Gamma}}_k)||_F^2/n_k \leq 16\rho_k^2 s_2(8/\mu_{\min})^2, \quad ||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}||_1 \leq 8\rho_k s_2(8/\mu_{\min})^2. \tag{57}$$

Since $\mu_{\min} \in (0, \infty)$, there exists $L = O(1)$ such that $(8/\mu_{\min})^2 \leq L$, and by considering $\rho_k \asymp \sqrt{\log(pq)/n_k}$, the result of (54) follows. □

Lemma 5. Consider the regression model (2) where $\mathbf{X}_k$ satisfies assumptions (A1) and (A2). Suppose that the eigenvalues of $\mathbf{\Sigma}$, the covariance matrix of the noise, are uniformly bounded. On the event $\mathcal{F}_k(n_k, p, q)$ with regularization parameter $\rho_k \asymp \sqrt{\log(pq)/n_k}$, we have

$$||\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}||_\infty = O_p\big(\max\{\sqrt{\log(p)/n_k}, s_2 \log(pq)/n_k\}\big), \tag{58}$$

and under the sparsity condition $s_2 = o\big(n_k^{\pi_1}/(\log(q)\log(pq))\big)$, $0 < \pi_1 \leq 1/2$, we get $||\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}||_\infty = o_p(1)$.

Proof. Recalling that $\xi_k = \mathbf{Y}_k - \mathbf{X}_k\mathbf{\Gamma}$, by adding and subtracting $\xi_k^\top \xi_k/n_k$ to $\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}$, we get

$$||\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}||_\infty \leq ||\xi_k^\top \xi_k/n_k - \mathbf{\Sigma}||_\infty + ||\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \xi_k^\top \xi_k/n_k||_\infty. \tag{59}$$

Under uniformly boundedness of the eigenvalues of $\mathbf{\Sigma}$, using Lemma 2 of Lam and Fan (2009),

$$||\xi_k^\top \xi_k/n_k - \mathbf{\Sigma}||_\infty = O_p(\sqrt{\log(p)/n_k}). \tag{60}$$

To find an upper bound on the second term of (59), by adding and subtracting $\mathbf{X}_k\mathbf{\Gamma}$ to $\mathbf{Y}_k - \mathbf{X}_k\hat{\mathbf{\Gamma}}_k$, we get

$$\begin{aligned}
&||\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \xi_k^\top \xi_k/n_k||_\infty \\
&\quad = ||\big(\xi_k - \mathbf{X}_k(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})\big)^\top \big(\xi_k - \mathbf{X}_k(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})\big)/n_k - \xi_k^\top \xi_k/n_k||_\infty \\
&\quad \leq 2||(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})^\top \mathbf{X}_k^\top \xi_k/n_k||_\infty + ||(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})^\top \mathbf{X}_k^\top \mathbf{X}_k(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})/n_k||_\infty.
\end{aligned} \tag{61}$$

Using Lemma 4, on the event $\mathcal{F}_k(n_k, p, q)$,

$$2||(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})^\top \mathbf{X}_k^\top \xi_k/n_k||_\infty \leq 2||\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma}||_1 ||\mathbf{X}_k^\top \xi_k||_\infty/n_k = O_p(s_2 \log(pq)/n_k). \tag{62}$$

On the other hand,

$$||(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})^\top \mathbf{X}_k^\top \mathbf{X}_k(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})/n_k||_\infty \leq ||\mathbf{X}_k(\hat{\mathbf{\Gamma}}_k - \mathbf{\Gamma})||_F^2/n_k = O_p(s_2 \log(pq)/n_k), \tag{63}$$

where the last equality is due to the fact that in (57), $\mu_{\min} \in (0, \infty)$ and $\rho_k \asymp \sqrt{\log(pq)/n_k}$. Substituting (62) and (63) in (61) and then substituting (61) and (60) in (59), the result in (58) follows. □

Lemma 6. Consider the regression model (2) where $\mathbf{X}_k$ satisfies assumptions (A1) and (A2), and consider the maximum row sparsity of $\mathbf{Q}^{-1}$ as $d_2 = o(\sqrt{n_\dagger}/\log(q))$. Moreover, consider the coefficient matrix $\mathbf{\Gamma}$ with sparsity condition $s_2 = o\big(n_\dagger^{\pi_1}/(\log(pq)\log(q))\big)$, $0 < \pi_1 \leq 1/2$. Suppose that $\hat{\mathbf{\Gamma}}_k^d$, $k = 1, \ldots, K$, is the k-th debiased estimator in (5) with tuning parameter $\rho_k \asymp \sqrt{\log(pq)/n_k}$ and $\tilde{\mathbf{\Gamma}}_{\text{owAvg}}$ is the pooled estimator in (17). Let $n_k/n \to c_k \in (0,1)$ as n_k grows, such that $\lim_{K\to\infty}\sum_{k=1}^K c_k = 1$, and consider the remainder term $\mathbf{R}_{a,\mathbf{\Gamma}}$ defined in (18). Then it follows that

$$|\mathbf{R}_{a,\mathbf{\Gamma}}| = O_p\big(K s_2 \sqrt{\log(pq)\log(q)/n}\big), \tag{64}$$

and if K grows at the rate $K = O\big(n^{\pi_4}/(\sqrt{\log(pq)\log(q)}\max\{s_2, \sqrt{d_2}\})\big)$, $0 < \pi_4 < 1/4$, we have $|\mathbf{R}_{a,\mathbf{\Gamma}}| = o_p(1)$.

Proof. First note that using Lemma 5, $|[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} - \mathbf{\Sigma}]_{aa}| = O_p\big(\max\{\sqrt{\log(p)/n_k}, s_2\log(pq)/n_k\}\big)$, where we recall that $\mathbf{A}_{aa}$ is the a-th diagonal element of an arbitrary matrix $\mathbf{A}$. Moreover, under the sparsity condition $d_2 = o(\sqrt{n_\dagger}/\log(q))$, with tuning parameter $\tilde{\rho}_{j,k} \asymp \sqrt{\log(q)/n_k}$ in the nodewise Lasso regression (28), using Lemma 5.4 of van de Geer et al. (2014), we get $|[\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top - \mathbf{Q}^{-1}]_{aa}| = O_p(\sqrt{d_2\log(q)/n_k})$. As such,

$$\begin{aligned}\big|[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa} - [\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}\big| = O_p\big(\max\{&\sqrt{d_2\log(q)/n_k},\\ &\sqrt{\log(p)/n_k}, s_2\log(pq)/n_k\}\big),\end{aligned} \tag{65}$$

where under the sparsity conditions $d_2 = o(\sqrt{n_\dagger}/\log(q))$ and $s_2 = o\big(n_\dagger^{\pi_1}/(\log(pq)\log(q))\big)$, $0 < \pi_1 \leq 1/2$, we get

$$\big|[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa} - [\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}\big| = o_p(1).$$

As such, $[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}$ is a consistent estimator for $[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}$. Using the continuous map $g(x) = 1/x$, we get $[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa} \xrightarrow{p} 1$ and since $n_k/n \to c_k$ then $(n_k/n)\times[\mathbf{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}/[\hat{\mathbf{\Sigma}}_{k,\hat{\mathbf{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa} \xrightarrow{p} c_k$ as $n_k \to \infty$. Using the dominated convergence theorem and the assump-

tion $\lim_{K\to\infty}\sum_{k=1}^{K} c_k = 1$, we get

$$\frac{\sum_{k=1}^{K} n_k/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}{\sum_{k=1}^{K} n_k/[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}} = \sum_{k=1}^{K}\left\{\frac{n_k/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}{\sum_{k=1}^{K} n_k/[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}\right\}$$
$$= \sum_{k=1}^{K}\left\{\frac{n_k}{n}\times\frac{[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\right\}\xrightarrow{p} 1, \tag{66}$$

as $K\to\infty$ and $n_k\to\infty$, $k=1,\ldots,K$. Again, by considering the continuous map $g(x)=1/\sqrt{x}$, the sequence $\sqrt{\frac{\sum_{k=1}^{K} n_k/[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}{\sum_{k=1}^{K} n_k/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}$ converges in probability to 1 as K and n_k, $k=1,\ldots,K$, grow. As such, due to the definition of $\mathbf{R}_{a,\boldsymbol{\Gamma}}$, it is enough to show the bound (64) for $\sum_{k=1}^{K}\frac{\sqrt{n_k[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}{\sqrt{n}[\hat{\boldsymbol{\Sigma}}_k\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\mathbf{R}_{a,k,\boldsymbol{\Gamma}}$, with $\mathbf{R}_{a,k,\boldsymbol{\Gamma}}$ the a-th element, $a=1,\ldots,qp$, of $\mathbf{R}_{k,\boldsymbol{\Gamma}}$ from (5).

We first show that $1/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa} = O_p(1)$. Note that $[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}$ is nothing else than the a-th element of the outer product of the diagonal elements of $\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}$ and $\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top$. Thus, it is just needed to find a lower bound for the diagonal elements of these two matrices. By construction, the diagonal elements of $\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}$ are always positive. Combining Theorem 2.2 of van de Geer et al. (2014) and the reversed triangle inequality, for each $a\in\{1,\ldots,q\}$ we get

$$|\mathbf{Q}_{aa}^{-1}| - |[\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top]_{aa}| \leq |\mathbf{Q}_{aa}^{-1} - [\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top]_{aa}| = o_p(1). \tag{67}$$

Since $\mathbf{Q}^{-1}$ is positive definite, using assumption (A2), we get $\mathbf{Q}_{aa}^{-1}\geq\Lambda_{\min}(\mathbf{Q}^{-1})$ $> 1/\Lambda_1$. As such, for sufficiently large n_k, using (67) we have that $|[\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top]_{aa}| > 0$. Thus, for every $a,b,c\in\{1,\ldots,q\}$, there exists a bound J_k such that

$$\frac{1}{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}} \leq \frac{1}{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}]_{bb}|[\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top]_{cc}|} \leq \frac{1}{J_k} = O_p(1). \tag{68}$$

By the fact that $[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}$ does not grow with n, using Theorem 1 and combining it with (68), we have that

$$\Big|\sum_{k=1}^{K}\frac{\sqrt{n_k[\boldsymbol{\Sigma}\otimes\mathbf{Q}^{-1}]_{aa}}}{\sqrt{n}[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k}\otimes(\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}\mathbf{R}_{a,k,\boldsymbol{\Gamma}}\Big| \leq \sum_{k=1}^{K}\sqrt{n_k/n}\big|\mathbf{R}_{a,k,\boldsymbol{\Gamma}}\big|\times O_p(1)$$
$$\leq O_p(Ks_2\sqrt{\log(q)\log(pq)/n}).$$

Considering the sparsity conditions $s_2 = o\big(n_\dagger^{\pi_1}/(\log(pq)\log(q))\big)$, $0 < \pi_1 \leq 1/2$, $d_2 = o(\sqrt{n_\dagger}/\log(q))$ and $K = O\big(n^{\pi_4}/(\sqrt{\log(pq)\log(q)}\max\{s_2, \sqrt{d_2}\})\big)$, $0 < \pi_4 < 1/4$, the $o_p(1)$ result follows immediately. □

Lemma 7. Under the assumptions of Theorem 4, by considering

$$\zeta_a = \sum_{k=1}^{K} \sqrt{\frac{n_k[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}{n[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}},$$

where $\mathbf{T}_{a,k}$ is the a-th element of $\mathbf{T}_k$ from (5), and

$$\zeta_a' = \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}},$$

we have $|\zeta_a - \zeta_a'| \xrightarrow{p} 0$ on the event $\mathcal{F}_k(n_k, p, q)$, as $K \to \infty$ and $n_k \to \infty$, $k = 1, \ldots, K$.

Proof. Denoting the sequence $\zeta_a'' = \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} \times \frac{\mathbf{T}_{a,k}}{\sqrt{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa}}}$, we show that $|\zeta_a - \zeta_a''| \xrightarrow{p} 0$ and $|\zeta_a'' - \zeta_a'| \xrightarrow{p} 0$, and then automatically, $|\zeta_a - \zeta_a'| \xrightarrow{p} 0$. Using (68), $1/[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top)]_{aa} = O_p(1)$. Thus,

$$\big|\zeta_a - \zeta_a''\big| \leq \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} \Big| \sqrt{[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}} - \sqrt{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes \mathbf{M}_k\mathbf{C}_k\mathbf{M}_k^\top]_{aa}} \Big| \times \big|\mathbf{T}_{a,k}\big|.$$

Moreover, using (65) we get

$$\begin{aligned}\big|\zeta_a - \zeta_a''\big| \leq \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} \big|\mathbf{T}_{a,k}\big| \times O_p\big(\max\{ &\sqrt{d_2 \log(q)/n_k}, \sqrt{\log(p)/n_k}, \\ &s_2 \log(pq)/n_k\}\big). \end{aligned} \tag{69}$$

On the other hand,

$$\mathbf{T}_k = \mathrm{vec}(\mathbf{M}_k\mathbf{X}_k^\top\xi_k/\sqrt{n_k}) = (\mathbf{I}_p \otimes \mathbf{M}_k)\mathrm{vec}(\mathbf{X}_k^\top\xi_k)/\sqrt{n_k},$$

and working on the event $\mathcal{F}_k(n_k, p, q)$, by considering $\rho_k \asymp \sqrt{\log(pq)/n_k}$,

$$||\mathbf{T}_k||_\infty \leq \sqrt{n_k} O_p(\sqrt{\log(pq)/n_k}) |||\mathbf{I}_p \otimes \mathbf{M}_k|||_\infty. \tag{70}$$

By the same argument as in the proof of Lemma 5.4 from van de Geer et al. (2014),

$$|||\mathbf{I}_p \otimes \mathbf{M}_k|||_\infty = |||\mathbf{M}_k|||_\infty = \max_{j\in\{1,\ldots,q\}} ||\mathbf{M}_{j,k}||_1 = O_p(\sqrt{d_2}),$$

where $\mathbf{M}_{j,k}$ is the j-th row of $\mathbf{M}_k$. As such, in (70), we get $||\mathbf{T}_k||_\infty \leq O_p(\sqrt{d_2 \log(pq)})$. Substituting this result in (69), we have

$$\begin{aligned}
\big|\zeta_a - \zeta_a''\big| &\leq \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} O_p\big(\sqrt{d_2 \log(pq)} \times \max\{\sqrt{d_2 \log(p)/n_k}, \sqrt{\log(p)/n_k}, \\
&\qquad s_2 \log(pq)/n_k\}\big) \\
&\leq O_p\big(K\sqrt{d_2 \log(pq)/n} \times \max\{\sqrt{d_2 \log(q)}, \sqrt{\log(p)}, \\
&\qquad s_2 \log(pq)/\sqrt{n_\dagger}\}\big),
\end{aligned}$$

where by considering the sparsity conditions $s_2 = o\big(n_\dagger^{\pi_1}/(\log(pq)\log(q))\big)$, $0 < \pi_1 \leq 1/2$, $d_2 = o(\sqrt{n_\dagger}/\log(q))$ and $K = O\big(n^{\pi_4}/(\sqrt{\log(pq)\log(q)} \times \max\{s_2, \sqrt{d_2}\})\big)$, $0 < \pi_4 < 1/4$, the $o_p(1)$ result follows immediately.

Now, to show $\big|\zeta_a'' - \zeta_a'\big| \xrightarrow{p} 0$, by the same argument,

$$\begin{aligned}
\big|\zeta_a'' - \zeta_a'\big| &\leq \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} \Big|\sqrt{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} - \sqrt{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}\Big| \\
&\qquad \times \big|\mathbf{T}_{a,k}\big| \\
&\leq \sum_{k=1}^{K} \sqrt{\frac{n_k}{n}} \Big\{\Big|\sqrt{[\boldsymbol{\Sigma} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}} - \sqrt{[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}}\Big| \\
&\qquad + \Big|\sqrt{[\boldsymbol{\Sigma} \otimes \mathbf{Q}^{-1}]_{aa}} - \sqrt{[\hat{\boldsymbol{\Sigma}}_{k,\hat{\boldsymbol{\Gamma}}_k} \otimes (\mathbf{M}_k \mathbf{C}_k \mathbf{M}_k^\top)]_{aa}}\Big|\Big\} \big|\mathbf{T}_{a,k}\big|.
\end{aligned}$$

Similarly to the proof of $|\zeta_a - \zeta_a''| \xrightarrow{p} 0$, by considering the mentioned sparsity conditions, we conclude that $\big|\zeta_a'' - \zeta_a'\big| \xrightarrow{p} 0$. □

Lemma 8. Consider the regression model (2), and suppose that assumptions (B1)-(B2), (C1)-(C3) from Appendix A.3 and (46) from Appendix B hold. Consider the coefficient matrix $\boldsymbol{\Gamma}$ with sparsity condition $s_2 = o(n_\dagger^{\pi_3}/\log(pq))$, $0 < \pi_3 \leq 1/6$, and the precision matrix $\boldsymbol{\Theta}$ with maximum

node degree $d_1^{3/2} = o(\sqrt{n_\dagger}/\log(p))$. Suppose that $\hat{\mathbf{\Theta}}_k^d$, $k = 1, \ldots, K$, is the k-th debiased estimator in (10) with tuning parameter $\lambda_k \asymp \sqrt{\log(p)/n_k}$ and $\tilde{\mathbf{\Theta}}_{\text{owAvg}}$ is the pooled estimator in (19). Let $n_k/n \to c_k \in (0,1)$ as n_k grows, such that $\lim_{K\to\infty} \sum_{k=1}^K c_k = 1$, and consider the remainder term $\mathbf{R}_{ab,\mathbf{\Theta}}$ defined in (20). Then it follows that

$$|\mathbf{R}_{ab,\mathbf{\Theta}}| = O_p\bigg(\frac{K}{\sqrt{n}} \max\big\{d_1^{3/2}\log(p), d_1^2(\log(p))^{3/2}/\sqrt{n_\dagger}, d_1 s_2 \log(pq)\big\}\bigg), \quad (71)$$

and if K grows as $K = O\big(n^{\pi_7}/(d_1\log(p))\big)$, $0 < \pi_7 \leq 1/3$, we have $|\mathbf{R}_{ab,\mathbf{\Theta}}| = o_p(1)$.

Proof. By a similar argument as in the proof of Lemma 6 and using Remark 4, the sequence $\sqrt{\frac{\sum_{k=1}^K n_k/\sigma_{ab}^2}{\sum_{k=1}^K n_k/\hat{\sigma}_{ab,k}^2}}$ converges in probability to 1 as K and n_k, $k = 1, \ldots, K$, grow. As such, we just need to show the boundedness of $\sum_{k=1}^K \frac{\sqrt{n_k}\sigma_{ab}}{\sqrt{n}\hat{\sigma}_{ab,k}^2}\mathbf{R}_{ab,k,\mathbf{\Theta}}$. Due to the positive definiteness of the graphical Lasso estimator defined in (8), there exists a positive constant L_k such that with high probability $\hat{\sigma}_{ab,k}^2 \geq \Lambda_{\min}^2(\hat{\mathbf{\Theta}}_k) > L_k$, where $\Lambda_{\min}(\hat{\mathbf{\Theta}}_k)$ is the minimum eigenvalue of $\hat{\mathbf{\Theta}}_k$ and then the term $1/\hat{\sigma}_{ab,k}^2 = O_p(1)$. Moreover, using (12),

$$\Big|\sum_{k=1}^K \frac{\sqrt{n_k}\sigma_{ab}}{\sqrt{n}\hat{\sigma}_{ab,k}^2}\mathbf{R}_{ab,k,\mathbf{\Theta}}\Big| \leq \sum_{k=1}^K \frac{1}{\sqrt{n}} O_p\big(\max\{d_1^{3/2}\log(p), d_1^2(\log(p))^{3/2}/\sqrt{n_\dagger}, d_1 s_2 \log(pq)\}\big),$$

and the result in (71) follows directly. By considering the mentioned sparsity and $K = O\big(n^{\pi_7}/(d_1\log(p))\big)$, $0 < \pi_7 \leq 1/3$, one can reach the $o_p(1)$ rate. □

C Extra simulation results

Table 4: Average and standard deviation (between parentheses) of the Frobenius norm over 500 repetitions on the active and non-active sets, for the proposed estimators and different competitors, when $n = 25000$.

			Active set				Non-active set			
					K				K	
			1	5	10	20	1	5	10	20
$\boldsymbol{\Theta}$	Debiased	Full	1.00				6.50			
			(.01)				(.01)			
		owAvg		1.03	1.08	1.18		6.72	7.15	7.79
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		Top1		1.39	1.46	1.19		9.12	9.60	7.85
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		sAvg		1.59	2.88	3.95		10.27	16.91	20.15
				(.01)	(.02)	(.03)		(.03)	(.02)	(.03)
		wAvg		1.04	1.25	1.81		6.80	8.00	10.10
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
	Sparse	SFull	2.74				.35			
			(.01)				(.00)			
		STop1		2.82	2.84	2.88		1.31	1.56	2.28
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		SsAvg		2.54	2.29	2.04		4.29	7.37	8.83
				(.01)	(.01)	(.01)		(.01)	(.01)	(.01)
		SwAvg		2.59	2.45	2.22		1.93	2.77	4.01
				(.01)	(.01)	(.01)		(.00)	(.00)	(.01)
$\boldsymbol{\Gamma}$	Debiased	Full	1.55				15.43			
			(.01)				(.02)			
		owAvg		1.62	1.74	1.55		16.18	17.30	14.28
				(.02)	(.02)	(.01)		(.02)	(.02)	(.02)
		Top1		2.21	2.34	2.67		22.08	23.33	26.65
				(.02)	(.02)	(.03)		(.03)	(.03)	(.03)
		sAvg		2.42	3.91	7.09		24.17	39.04	70.18
				(.02)	(.04)	(1.45)		(.03)	(.06)	(14.36)
		wAvg		1.63	1.90	3.38		16.30	18.93	33.51
				(.02)	(.02)	(.61)		(.02)	(.02)	(6.03)
	Sparse	SFull	4.79				1.82			
			(.00)				(.00)			
		STop1		4.80	4.80	4.80		1.91	1.93	1.99
				(.00)	(.00)	(.00)		(.00)	(.00)	(.01)
		SsAvg		4.67	4.23	4.17		2.03	2.82	3.11
				(.12)	(.02)	(.02)		(.02)	(.01)	(.01)
		SwAvg		4.75	4.59	4.51		1.82	1.84	1.98
				(.04)	(.01)	(.01)		(.02)	(.00)	(.00)

Table 5: Average coverage probability and average length of the confidence intervals over 500 repetitions for the proposed estimators and different competitors, when $n = 25000$.

			Avg.Cov				Avg.Len			
			K				K			
			1	5	10	20	1	5	10	20
$\boldsymbol{\Theta}$	Active set	Full	.94				.02			
		owAvg		.94	.94	.95		.02	.03	.03
		Top1		.94	.94	.94		.03	.04	.03
		sAvg		.92	.89	.90		.04	.06	.08
		wAvg		.95	.98	.98		.03	.04	.06
	Non-active set	Full	.95				.02			
		owAvg		.95	.96	.96		.02	.03	.03
		Top1		.96	.96	.95		.03	.04	.03
		sAvg		.94	.94	.97		.04	.06	.08
		wAvg		.97	.98	.98		.03	.04	.06
$\boldsymbol{\Gamma}$	Active set	Full	.95				.08			
		owAvg		.95	.95	.92		.08	.09	.07
		Top1		.95	.95	.95		.11	.12	.14
		sAvg		.94	.92	.90		.12	.17	.27
		wAvg		.96	.97	.98		.09	.12	.20
	Non-active set	Full	.95				.08			
		owAvg		.95	.95	.94		.08	.09	.07
		Top1		.95	.95	.95		.11	.12	.14
		sAvg		.94	.92	.90		.12	.17	.27
		wAvg		.96	.97	.98		.09	.12	.20

References

Akbani, R., K. C. Akdemir, B. A. Aksoy, M. Albert, A. Ally, S. B. Amin, H. Arachchi, A. Arora, J. T. Auman, B. Ayala, et al. (2015). Genomic classification of cutaneous melanoma. *Cell 161*(7), 1681–1696.

Banerjee, O., L. E. Ghaoui, and A. d'Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research 9*(3), 485–516.

Battey, H., J. Fan, H. Liu, J. Lu, and Z. Zhu (2018). Distributed testing and estimation under sparse high dimensional models. *The Annals of Statistics 46*(3), 1352–1382.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer.

Cai, T., W. Liu, and X. Luo (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*(494), 594–607.

Cai, T., W. Liu, and H. Zhou (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics 44*(2), 455–488.

Cai, T. T., H. Li, W. Liu, and J. Xie (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika 100*(1), 139–156.

Cancer Genome Atlas Research Network (2017). Integrated genomic and molecular characterization of cervical cancer. *Nature 543*(7645), 378–384.

Chen, M., Z. Ren, H. Zhao, and H. Zhou (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *Journal of the American Statistical Association 111*(513), 394–406.

Chen, X. and M. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 1655–1684.

Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Gut, A. (2005). *Probability: a graduate course*, Volume 200. Springer.

Huo, X. and S. Cao (2019). Aggregated inference. *Wiley Interdisciplinary Reviews: Computational Statistics 11*(1), e1451.

Jankova, J. and S. van de Geer (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics 9*(1), 1205–1229.

Javanmard, A. and A. Montanari (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics 46*(6A), 2593–2622.

Jordan, M. I., J. D. Lee, and Y. Yang (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Kallenberg, O. (1997). *Foundations of modern probability* (2nd ed.). Springer.

Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics 37*(6B), 4254–4278.

Lee, J. D., Q. Liu, Y. Sun, and J. E. Taylor (2017). Communication-efficient sparse regression. *The Journal of Machine Learning Research 18*(1), 115–144.

Liu, J., T. Lichtenberg, K. A. Hoadley, L. M. Poisson, A. J. Lazar, A. D. Cherniack, A. J. Kovatich, C. C. Benz, D. A. Levine, A. V. Lee, et al. (2018). An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell 173*(2), 400–416.

Loh, P.-L. and X. L. Tan (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics 12*(1), 1429–1467.

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Nezakati, E. and E. Pircalabelu (2023). Unbalanced distributed estimation and inference for the precision matrix in Gaussian graphical models. *Statistics and Computing 33*, 1–14.

Obozinski, G., M. J. Wainwright, and M. I. Jordan (2011). Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics 39*(1), 1–47.

Peng, J., J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. Wang (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics 4*(1), 53–77.

Raskutti, G., M. J. Wainwright, and B. Yu (2010). Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research 11*, 2241–2259.

Ravikumar, P., M. J. Wainwright, G. Raskutti, and B. Yu (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics 5*, 935–980.

Rothman, A. J., E. Levina, and J. Zhu (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics 19*(4), 947–962.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics 42*(3), 1166–1202.

Wang, J. (2015). Joint estimation of sparse multivariate regression and conditional graphical models. *Statistica Sinica*, 831–851.

Wang, X., L. Qin, H. Zhang, Y. Zhang, L. Hsu, and P. Wang (2015). A regularized multivariate regression approach for eqtl analysis. *Statistics in biosciences 7*(1), 129–146.

Yin, J. and H. Li (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics 5*(4), 2630–2650.

Yin, J. and H. Li (2013). Adjusting for high-dimensional covariates in sparse precision matrix estimation by ℓ_1-penalization. *Journal of Multivariate Analysis 116*, 365–381.

Yuan, M. and Y. Lin (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika 94*(1), 19–35.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(1), 217–242.